May 13-17, 2019
San Francisco, CA, USA

Association for Computing Machinery

Advancing Computing as a Science & Profession

THE WEB CONFERENCE

The Web Conference 2019

Companion of

The World Wide Web Conference WWW 2019

Editors:

Sihem Amer-Yahia (CRNS & Université Grenoble Alps, France)
Mohammad Madian (Google, USA)
Ashish Goel (Stanford University, USA)
Geert-Jan Houben (TU Delft, The Netherlands)
Kristina Lerman (Information Sciences Institute, USA)
Julian McAuley(University of California at San Diego, USA)
Ricardo Baeza-Yates (NTENT & Northeastern University, USA)
Leila Zia (Wikimedia Foundation, USA)

THE WEB CONFERENCE

IW3C2

Copyright © 2019 by IW3C2 (International World Wide Web Conference Committee). The papers are published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution. In case of republication, reuse, etc., the following attribution should be used: "Published in WWW2019 Proceedings © 2019 International World Wide Web Conference Committee, published under Creative Commons CC By 4.0 License."

ISBN: 978-1-4503-6675-5

Printed in the USA

The Web Conference 2019
Conference Chairs' Welcome

It is our great pleasure to welcome you to *The Web Conference 2019*. The Web Conference is the premier venue focused on understanding the current state and the evolution of the Web through the lens of computer science, computational social science, economics, policy, and many other disciplines. The 2019 edition of the conference is a reflection point as we celebrate the 30th anniversary of the Web.

The technical program of this year's edition of the conference consists of three keynotes (Jeff Dean, Lawrence Lessig and Claire Wardle), one panel to celebrate the 30 years of the Web, 12 tracks for refereed research papers (including full papers and short papers), and a demonstrations track that are published in the main volume of the proceedings. The companion volume of the proceedings includes 23 workshops, most of the 22 tutorials, and the Ph.D. Consortium. We also have several other co-located events and special tracks: W4A, BIG, The Web of Health, The Open Web, Building an Ethical Web, and W3C.

We are thankful to all the people that helped us organizing the conference and that are listed in the next section. Without them, this would have never happened. Special thanks to all our sponsors, as they make the quality difference: Microsoft (Diamond); Amazon, Bloomberg and Google (Gold); Criteo AI Lab, CISCO, NTENT, Spotify, Yahoo Research and Wikimedia Foundation (Silver); Baidu, DiDi, eBay, Facebook, LinkedIn, Megagon Labs, Mix, Mozilla, Netflix Research, Northeastern University, Pinterest, Quora, Visa Research and Walmart Labs (Bronze); as well as Airbnb, Letgo, Moore Foundation and Webcastor.

Finally, this conference would have not been possible if together with Ashish Goel and Ben Steichen we did not start Web4Good, a California based non-profit company that undersigned the event with IW3C2.

So we hope that you enjoy this week as well as San Francisco!

Ricardo Baeza-Yates
General Co-Chair
NTENT & Northeastern University at SV
USA

Leila Zia
General Co-Chair
Wikimedia Foundation
USA

The Web Conference 2019
Companion Volume Preface

The companion volume of this year's edition of The Web Conference include all the workshops proceedings, summaries of most tutorials, and the PhD Symposium contributions.

The Workshop Track received 45 submissions, which ultimately culminated in 23 workshops (among which many of the original proposals were combined). Workshops were presented either as half- or full-day forums during the first two days of the conference, and included presentations, poster sessions, and panel discussions. Each workshop submission was given 10 pages in this volume.

Thirty-five submissions were made to the Tutorials track, among which 22 were accepted for inclusion in the conference. A summary of each tutorial of at most 2 pages is included in this volume.

The PhD Symposium attracted 18 submissions, among which 13 were accepted for presentation at the conference. The accepted submissions were presented during a dedicated session at the conference, geared toward giving PhD students feedback prior to the completion of their dissertation, where they were mentored by Ceren Budak, Ben Steichen, and Markus Strohmaier. They also had the opportunity to show their work in one of the posters session. Each PhD Symposium submission was given 5 pages in this volume.

Finally, we thank the conference chairs, Leila and Ricardo, that helped in the selection process of tutorials and workshops, as well as in the presentation format of the PhD symposium.

Sihem Amer-Yahia
Workshops Co-Chair
CRNS & Université Grenoble Alps, France

Mohammad Madian
Workshops Co-Chair
Google, USA

Ashish Goel
Tutorials Chair
Stanford University, USA

Geert-Jan Houben
PhD Symposium Co-Chair
TU Delft, The Netherlands

Kristina Lerman
PhD Symposium Co-Chair
Information Sciences Institute, USA

Julian McAuley
Proceedings Chair
University of California at San Diego, USA

Table of Contents

The Web Conference Organization xviii

Alternate tracks program committee members -
The Web Conference Sponsors & Partners xvix

PhD Symposium Chairs' Welcome

- **Incorporating System-Level Objectives into Recommender Systems** 2
  Himan Abdollahpouri *(University of Colorado Boulder, USA)*
- **Identification of the Source(s) of Misinformation Propagation Utilizing Identifying Codes** 7
  Kaustav Basu *(Arizona State University, USA)*
- **Representation Learning for Recommender Systems with Application to the Scientific Literature** 12
  Robin Brochier *(Université de Lyon, Lyon 2, ERIC EA 3083, France)*
- **Interactive Visualisation Techniques for the Web of Data** 17
  Marie Destandau *(INRIA, France)*
- **Cores and Other Dense Structures in Complex Networks** 22
  Edoardo Galimberti *(University of Turin & ISI Foundation, Italy)*
- **Extracting actionable information from Security Forums** 27
  Joobin Gharibshah, Michalis Faloutsos *(University of California, Riverside, USA)*
- **Dynamics of Misinformation Cascades** 33
  Jasser Jasser *(University of Central Florida, USA)*
- **ClickGraph: Web Page Embedding using Clickstream Data for Multitask Learning** 37
  Porter Jenkins *(The Pennsylvania State University, USA)*
- **Finding Themselves between Home and Host Cultures** 42
  Lin Malone *(University of Western Australia, Australia)*
- **Data-Driven Edu Chatbots** 46
  Donya Rooein *(Politecnico di Milano, Italy)*
- **Towards Better Entity Linking Evaluation** 50
  Henry Rosales-Méndez *(University of Chile, Chile)*
- **Velocity on the Web - a PhD Symposium** 56
  Riccardo Tommasini *(Politecnico di Milano, Italy)*
- **Robust Detection of Cyberbullying in Social Media** 61
  Mengfan Yao *(Univeristy at Albany, USA)*

Attention Workshop Chairs' Welcome

- **Scalable Eye Tracking for Mobile Devices** 69
  Michael Xuelin Huang *(Google)*
- **Lessons from neuroscience** 70
  Laurent Itti *(USC)*

5th AW4City 2019 Chairs' Welcome

- **Persuasive Technologies for Sustainable Smart Cities: The Case of Urban Mobility** 73
  Evagelia Anagnostopoulou, Babis Magoutas, Efthimios Bothos, Gregoris Mentzas *(National Technical Univ. of Athens)*
- **Mobility as a Service** 83
  Rob Christiaanse *(Univ. of Technology of Delft)*
- **Supporting Urban Air Mobility with Citizen Participatory Noise Sensing: A Concept** 93
  Hinnerk Eißfeldt *(DLR German Aerospace Center, Hamburg)*

- **Integrating Route Attractiveness Attributes into Tourist Trip Recommendations** 96
  Daniel Herzog, Sherjeel Sikander, Wolfgang Wörndl *(Technical Univ. of Munich)*
- **Hybrid Human Machine workflows for mobility management** 102
  Eddy Maddalena, Luis-Daniel Ibáñez, Elena Simperl, Richard Gomer *(Univ. of Southampton)*, Mattia Zeni, Donglei Song, Fausto Giunchiglia *(Univ. of Trento)*
- **Intelligent Traffic Congestion Classification System using Artificial Neural Network** 110
  Md Ashifuddin Mondal *(Narula Institute of Technology)*, Zeenat Rehena *(Aliah Univ.)*
- **Leveraging AI and Sensor Fabrics to Evolve Smart City Solution Designs** 117
  William H. Money *(The Citadel Baker School of Business)*, Stephen Cohen *(Azure Global Microsoft Corp.)*
- **TOMI: A Framework for Smart Tourism on the Move Innovation** 123
  Vaia Moustaka, Athena Vakali *(Department of Informatics, Aristotle University, Thessaloniki, Greece)*, Nikos Zikos, Theoklis Tsirakidis *(Movvin PC, Thessaloniki, Greece)*, Leonidas G. Anthopoulos *(University of Thessaly Greece)*
- **Travel Time Estimation Accuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR** 130
  Daksh Shah, Aravinda Kumaran, Rijurekha Sen, Ponnurangam Kumaraguru *(IIIT Delhi)*
- **Data4UrbanMobility: Towards Holistic Data Analytics for Mobility Applications in Urban Regions** 137
  Nicolas Tempelmeier *(L3S Research Center, Leibniz Universität Hannover)*, Yannick Rietz *(PROJEKTIONISTEN GmbH)*, Iryna Lishchuk *(Institute for Legal Informatics, Leibniz Universität Hannover)*, Tina Kruegel *(Institute for Legal Informatics, Leibniz Universität Hannover)*, Olaf Mumm *(Institute for Sustainable Urbanism, TU Braunschweig)*, Vanessa Miriam Carlow *(Institute for Sustainable Urbanism, TU Braunschweig)*, Stefan Dietze *(GESIS - Leibniz Institute for the Social Sciences)*, Elena Demidova *(L3S Research Center, Leibniz Universität Hannover)*

CyberSafety 2019 Chairs' Welcome Message

- **Detect Me If You Can: Spam Bot Detection Using Inductive Representation Learning** 148
  Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, Christoph Meinel *(Hasso-Plattner Institute, University of Potsdam)*
- **Less is More: Semi-Supervised Causal Inference for Detecting Pathogenic Users in Social Media** 154
  Hamidreza Alvari, Elham Shaabani, Soumajyoti Sarkar, Ghazaleh Beigi *(Arizona State Univ.)*, Paulo Shakarian *(Arizona State University)*
- **Who Falls for Online Political Manipulation?** 162
  Adam Badawy, Kristina Lerman, Emilio Ferrara *(USC ISI)*
- **Robust Cyberbullying Detection with Causal Interpretation** 169
  Lu Cheng, Ruocheng Guo, Huan Liu *(Arizona State Univ.)*
- **An Empirical Study of Malicious Threads in Security Forums** 176
  Joobin Gharibshah *(University of California Riverside)*, Zhabiz Gharibshah *(Florida Atlantic University)*, Evangelos E. Papalexakis, Michalis Faloutsos *(University of California Riverside)*
- **Understanding Cyberbullying on Instagram and Ask.fm via Social Role Detection** 183
  Hsien-Te Kao, Shen Yan, Di Huang, Nathan Bartley, Homa Hosseinmardi, Emilio Ferrara *(University of Southern California)*
- **Domain Generation Algorithms detection through deep neural network and ensemble** 189
  Shuaiji Li, Tao Huang, Zhiwei Qin, Fanfang Zhang, Yinhong Chang *(DiDi Research America)*
- **Hi Doppelgänger : Towards Detecting Manipulation in News Comments** 197
  Jan Pennekamp, Martin Henze, Oliver Hohlfeld *(RWTH Aachen University)*, Andriy Panchenko *(Brandenburg University of Technology)*
- **Characterizing Activity on the Deep and Dark Web** 206
  Nazgol Tavabi, Nathan Bartley, Andres Abeliuk *(USC ISI)*, Sandeep Soni *(Georgia Institute of Technology)*, Emilio Ferrara, Kristina Lerman *(USC ISI)*
- **Bot Electioneering Volume: Visualizing Social Bot Activity During Elections** 214
  Kai-Cheng Yang, Pik-Mai Hui, Filippo Menczer *(Center for Complex Networks and Systems Research, Indiana University)*

- **Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web** 218
  Savvas Zannettou *(Cyprus Univ. of Technology)*, Tristan Caulfield, Emiliano De Cristofaro *(Univ. College London)*, Michael Sirivianos *(Cyprus University of Technology)*, Gianluca Stringhini *(Boston University)*, Jeremy Blackburn *(University of Alabama at Birmingham)*

Data Science for Social Good Chairs' Welcome

- **Problems and Opportunities of Working with a Telco's Large Data Sets of Mobile Data*** 229
  Leo Ferres, *(Universidad del Desarrollo, Chile)*
- **AI for Good at Scale: Challenges in Real Time Emergency Response** 230
  Alex Jaimes *(AI & Data Science division DataMinr, New York, USA)*
- **Creating Public Value by Democratizing the Ecosystem of Human Service Providers** 231
  Charalampos Chelmis, Mengfan Yao *(University at Albany, SUNY, Department of Computer Science)*
- **Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms** 237
  Ashok Deb *(USC Information Sciences Institute Marina del Rey, CA)*, Luca Luceri *(University of Applied Sciences and Arts of Southern Switzerland, and University of Bern Manno, Switzerland)*, Adam Badaway, Emilio Ferrara *(USC Information Sciences Institute Marina del Rey, CA)*
- **Human Values and Attitudes towards Vaccination in Social Media** 248
  Kyriaki Kalimeri *(ISI Foundation)*, Mariano G. Beiró *(FIUBA)*, Alessandra Urbinati *(ISI Foundation)*, Andrea Bonanomi, Alessandro Rosina *(Università Cattolica del sacro cuore)*, Ciro Cattuto *(ISI Foundation)*
- **Cross-lingual Perspectives about Crisis-Related Conversations on Twitter** 255
  Johnny Torres, Carmen Vaca
- **Detecting Areas of Potential High Prevalence of Chagas in Argentina** 262
  Antonio Vazquez Brust, Tomás Olego, Germán Rosati, Carolina Lang, Guillermo Bozzoli *(Fundación Bunge y Born)*, Diego Weinberg, Roberto Chuit *(Fundación Mundo Sano)*, Martin Minnoni, Carlos Sarraute *(Grandata Labs)*

DL4G-SDE 2019 Chairs' Welcome

- **Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain** 274
  Oghenemaro Anuyah, Ion Madrazo Azpiazu, Maria Soledad Pera *(Department of Computer Science, Boise State University, Boise, Idaho, USA)*
- **Link Prediction with Mutual Attention for Text-Attributed Networks** 283
  Robin Brochier *(Université de Lyon, Lyon 2, ERIC EA3083, Digital Scientific Research Technology)*, Adrien Guille, Julien Velcin *(Université de Lyon, Lyon 2, ERIC EA3083)*
- **Real-Time Streaming Graph Embedding Through Local Actions** 285
  Xi Liu, Ping-Chun Hsieh, Nick Duffield *(Texas A&M University)*, Rui Chen, Muhe Xie *(Samsung Research America)*, Xidao Wen *(University of Pittsburgh)*
- **Integrating Bayesian and Neural Networks for Discourse Coherence** 294
  Jinhua Peng, Zongyang Ma, Di Jiang, Hua Wu *(Baidu)*
- **Learning to Represent the Evolution of Dynamic Graphs with Recurrent Models** 301
  Aynaz Taheri *(UIC)*, Kevin Gimpel *(TTIC)*, Tanya Berger-Wolf *(UIC)*
- **Spectral Graph Complexity** 308
  Anton Tsitsulin *(Univ. of Bonn)*, Davide Mottin, Panagiotis Karras *(Aarhus Univ.)*, Alexander Bronstein *(Technion)*, Emmanuel Müller *(Univ. of Bonn)*
- **FdGars: Fraudster Detection via Graph Convolutional Networks in Online App Review System** 310
  Rui Wen *(Tencent Shenzhen, China)*, Jianyu Wang, Chunming *(Zhejiang University Hangzhou, China)*, Yu Huang, Jian Xion *(Tencent Shenzhen, China)*
- **Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings** 317
  Dongsheng Wang *(Univ. of Copenhagen, Denmark)*, Qiuchi Li *(Univ. of Padova, Italy)*, Lucas Chaves Lima, Jakob Grue Simonsen, Christina Lioma *(University of Copenhagen, Denmark)*

First International Workshop on e-Commerce and NLP (ECNLP) Chairs' Welcome

- **On the Relation Between Products' Relevance and Customers' Satisfaction in Voice Shopping** 326
  David Carmel *(Amazon, Israel)*
- **Shop your Right Size: A System for Recommending Sizes for Fashion products** 327
  G Mohammed Abdulla, Shreya Singh, Sumit Borar *(Myntra Designs, India)*
- **From Free-text User Reviews to Product Recommendation using Paragraph Vectors and Matrix Factorization** 335
  Georgios Alexandridis, Thanos Tagaris, Giorgos Siolas, Andreas Stafylopatis *(National Technical University of Athens)*
- **Exploration of Product Search Intents via Clustering of Query Clusters** 344
  Omar Alonso, Vasileios Kandylas, Serge-Eric Tremblay *(Microsoft)*
- **Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models** 348
  Anjan Goswami, Prasant Mohapatra *(UC David)*, Chengxiang Zhai *(UIUC)*
- **ProductQnA: Answering User Questions on E-Commerce Product Pages** 354
  Ashish Kulkarni, Kartik Mehta, Shweta Garg, Vidit Bansal, Nikhil Rasiwasia, Srinivasan Sengamedu *(India Machine Learning, Amazon)*
- **Behavior Matching between Different Domains based on Canonical Correlation Analysis** 361
  Yuan Lyu, Daichi Amagata, Takuya Maekawa, Takahiro Hara *(Osaka University)*, Hao Niu, Kei Yonekawa, Mori Kurokawa *(KDDI Research, Inc.)*
- **Neural Product Retrieval at Walmart.com** 367
  Alessandro Magnani, Feng Liu, Min Xie, Somnath Banerjee *(Walmart Labs Sunnyvale, United States)*
- **An Anatomy of a Lie:** 373
  Dina Pisarevskaya *(HSE)*, Boris Galitsky, Jay Taylor, Andrey Ozerov *(Oracle)*
- **The WDC Training Dataset and Gold Standard for Large-Scale Product Matching** 381
  Anna Primpeli, Ralph Peeters, Christian Bizer *(University of Mannheim)*
- **Towards language-agnostic alignment of product titles and descriptions: a neural approach** 387
  Daniel Stein *(eBay)*, Dimitar Shterionov, Andy Way *(ADAPT Centre)*
- **Product Phrase Extraction from e-Commerce Pages** 393
  Artem Vovk *(Google)*, Dmitrii Tochilkin *(Skolkovo Institute of Science and Technology)*, Pradyumna Narayana, Kazoo Sone, Sugato Basu *(Google)*
- **Preliminary Investigation of Alleviating User Cold-Start Problem in E-commerce with Deep Cross-Domain Recommender System** 398
  Hanxin Wang, Daichi Amagata, Takuya Maekawa, Takahiro Hara *(Osaka University)*, Hao Niu, Kei Yonekawa, Mori Kurokawa *(KDDI Research, Inc)*
- **Reference Product Search** 404
  Chu Wang, Lei Tang, Shujun Bian, Da Zhang, Zuohua Zhang, Yongning Wu *(Amazon.com)*
- **ProductNet: a Collection of High-Quality Datasets for Product Representation Learning** 411
  Chu Wang, Lei Tang, Yang Lu, Shujun Bian, Hirohisa Fujita, Da Zhang, Zuohua Zhang, Yongning Wu *(Amazon)*

Emoji2019 Workshop Chairs' Welcome

- **Decoding the New World Language: Analyzing the Popularity, Roles, and Utility of Emojis** 417
  Qiaozhu Mei *(Univ. of Michigan)*
- **Assessing the Meaning of Emojis for Emotional Awareness - A Pilot Study** 419
  Wesley Brants *(Eindhoven Univ. of Technology)*, Bonita Sharif *(Univ. of Nebraska - Lincoln)*, Alexander Serebrenik *(Eindhoven Univ. of Technology)*
- **When a Merperson is a Merman** 424
  Jennifer Daniel *(Google)*

- **Emoji Sequence Use in Enacting Personal Identity** 426
  Jing Ge *(UC Berkeley)*
- **Multimodal Emotion Classification** 439
  Anurag Illendula *(Indian Institute of Technology-Kharagpur)*, Amit Sheth *(Kno.e.sis Center, Wright State Univ.)*
- **Opico: A Study of Emoji-first Communication in a Mobile Social App** 450
  Sujay Khandekar, Joseph Higg, Yuanzhe Bian, Chae Won Ryu *(UIUC)*, Jerry O. Talton Iii *(Carta, Inc.)*, Ranjitha Kumar *(UIUC)*
- **Hanmoji: What Chinese Characters and Emoji Reveal About Each Other** 459
  Jennifer 8. Lee *(Emojination)*, Jason Li, An Xiao Mina *(Visual Culture Studies, The Civic Beat)*
- **Exploring (Dis-)Similarities in Emoji-Emotion Association on Twitter and Weibo** 461
  Mingyang Li, Sharath Guntuku *(Univ. of Pennsylvania)*, Vinit Jakhetiya *(Indian Institute of Technology)*, Lyle Ungar *(Univ. of Pennsylvania)*
- **Emoji Prediction for Hebrew Political Domain** 468
  Chaya Liebeskind, Shmuel Liebeskind *(Jerusalem College of Technology)*
- **Emojis in Sentence Processing: An Electrophysiological Approach** 478
  Benjamin Weissman *(UIUC)*

FATES on the WEB Chairs' Welcome

- **The Social Science of Privacy - Effects on Industry and Government Data Use** 483
  Frauke Kreuter *(Univ. of Maryland)*
- **Algorithms for Fair Team Formation in Online Labour Marketplaces** 484
  Giorgio Barnabò, Adriano Fazzone, Stefano Leonardi, Chris Schwiegelshohn *(Sapienza Univ. of Rome)*
- **Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification** 491
  Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, Lucy Vasserman *(Google)*
- **Collaborative Explanation of Deep Models with Limited Interaction for Trade Secret and Privacy Preservation** 501
  Josep Domingo-Ferrer, Cristina Pérez-Solà, Alberto Blanco-Justicia *(Universitat Rovira i Virgili, Dept. of Computer Science and Mathematics, UNESCO Chair in Data Privacy, CYBERCAT-Center for Cybersecurity, Research of Catalonia, Tarragona, Catalonia)*
- **Privacy-aware Linked Widgets** 508
  Javier D. Fernández *(Vienna University of Economics and Business, Complexity Science Hub Vienna)*, Fajar J. Ekaputra, Peb Ruswono Aryan *(Vienna University of Technology)*, Amr Azzam *(Vienna University of Economics and Business)*, Elmar Kiesling *(Vienna University of Technology)*
- **On Preserving Sensitive Information of Multiple Aspect Trajectories In-House** 515
  Spyros Giotakis, Nikos Pelekis *(Univ. of Piraeus)*
- **Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape** 523
  Jeanna Matthews *(Clarkson Univ.)*, Matt Goerzen *(Data and Society)*
- **Trust and Trustworthiness in Social Recommender Systems** 529
  Taha Hassan *(Virginia Tech)*
- **Empirical Analysis of Bias in Voice-based Personal Assistants** 533
  Lanna Lima, Vasco Furtado, Elizabeth Furtado *(Universidade de Fortaleza)*, Virgilio Almeida *(Universidade Federal de Minas Gerais)*
- **Managing Bias in AI** 539
  Drew Roselli *(ParallelM)*, Jeanna Matthews *(Clarkson Univ.)*, Nisha Talagala *(ParallelM)*
- **Uncovering Social Media Bots: a Transparency-focused Approach** 545
  Eric Ferreira Dos Santos, Danilo Carvalho, Livia Ruback, Jonice Oliveira *(Federal University of Rio de Janeiro, Rio de Janeiro, Rio de Janeiro)*
- **Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists** 553
  Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, Christo Wilson *(Northeastern Univ.)*
- **Unsupervised Topic Extraction from Privacy Policies** 563
  David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, Ittai Bar Siman Tov *(Bar-Ilan Univ.)*

- **Fairness in Social Influence Maximization** 569
  Ana-Andreea Stoica, Augustin Chaintreau *(Columbia Univ.)*
- **Hegemony in Social Media and the effect of recommendations** 575
  Ana-Andreea Stoica, Augustin Chaintreau *(Columbia University, USA)*
- **Privacy and Transparency within the 4IR: Two faces of the same coin** 581
  Bianca Teixeira, Daniel Schwabe *(PUC-Rio)*, Flavia Santoro *(UERJ)*, Fernanda Baião *(PUC-Rio)*, Maria Luiza Campos, Letícia Verona *(UERJ)*, Carlos Laufer, Simone Barbosa, Sérgio Lifschitz *(PUC-Rio)*, Rosa Costa *(UERJ)*
- **Achieving Differential Privacy and Fairness in Logistic Regression** 594
  Depeng Xu, Shuhan Yuan, Xintao Wu *(Univ. of Arkansas)*

Augmenting Intelligence with Bias-aware Humans-in-the-loop (HumBL2019) Chairs' Welcome

- **Are Two Heads Better Than One? An Exploration of Ambiguity in Crowd-Collected Language Decisions from the Phrase Detectives Game.** 603
  Jon Chamberlain *(Un of Essex)*
- **Identifying High Value Opportunities for Human in the Loop Lexicon Expansion** 604
  Alfredo Alba, Chad DeLuca, Anna Lisa Gentile, Daniel Gruhl, Linda Kato, Chris Kau, Petar Ristoski, Steve Welch *(IBM Research Almaden, CA, US)*
- **Task Oriented Data Exploration with Human-in-the-Loop. A Data Center Migration Use Case.** 610
  Alfredo Alba, Chad DeLuca, Anna Lisa Gentile, Daniel Gruhl, Linda Kato, Chris Kau, Petar Ristoski, Steve Welch *(IBM Research Almaden, USA)*
- **Uncovering Bias in Ad Feedback Data Analyses & Applications** 614
  Marc Bron *(Oracle)*, Ke Zhou *(University of Nottingham)*, Andy Haines *(Apple)*, Mounia Lalmas *(Spotify)*
- **Implicit Bias in Crowdsourced Knowledge Graphs** 624
  Gianluca Demartini *(University of Queensland, Australia)*
- **I'm Lonely. Who should I talk to?** 631
  Lauren Fratamico, Deb Roy *(Massachusetts Institute of Technology)*
- **Putting the Human in the Time Series Analytics Loop** 635
  Shima Imani, Sara Alaee, Eamonn Keogh *(University of California)*
- **Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering** 645
  Wenlong Sun, Sami Khenissi, Olfa Nasraoui *(University of Louisville)*, Patrick Shafto *(Rutgers University)*

Workshop on Hypermedia Multi-Agent Systems (HyperAgents) Chairs' Welcome

- **MAMS: Multi-Agent MicroServices** 655
  Rem W. Collier, Eoin O'Neill, David Lillis, Gregory O'Hare *(University College Dublin, Dublin, Ireland)*
- **Digital Companion for Industry** 663
  Mareike Kritzler, Jack Hodges, Dan Yu, Kimberly Garcia, Hemant Shukla, Florian Michahelles *(Siemens Corporate Technology Artificial)*
- **Towards Enabling Internet-Scale Context-as-a-Service: A Position Paper** 668
  Alexandru Sorici, Andrei Olaru, Adina Magda Florea *(Univ. Politehnica of Bucharest)*

The First International Workshop on Knowledge Graph Technology and Applications Chairs' Welcome

- **Invited Speakers – Morning Session** 674
- **Invited Speakers – Afternoon Session** 675
- **Panel: Knowledge Graph Industry Applications** 676
  Joshua Shinavier *(Uber)*, Kim Branson *(Genentech)*, Wei Zhang *(Alibaba)*, Shima Dastgheib *(numedii)*, Yuqing Gao *(Microsoft)*, Bogdan Arsintescu *(LinkedIn)*, Fatma Özcan *(IBM Almaden)*, Edgar Meij *(Bloomberg)*

- **Panel: Computational Methods about Knowledge Graph** 677
  Jie Tang *(Tsinghua University)*, Jure Leskovec *(Stanford)*, Kuansan Wang *(Microsoft)*, Deborah McGuinness *(RPI)*, Hassan Sawaf *(Amazon)*
- **Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolutional Network** 678
  Shumin Deng *(Zhejiang Univ, AZFT)*, Ningyu Zhang *(Alibaba Group, China)*, Wen Zhang *(Zhejiang University, AZFT Joint Lab for Knowledge Engine, China)*, Jiaoyan Chen *(University of Oxford, United Kingdom)*, Jeff Z. Pan *(University of Aberdeen, United Kingdom)*, Huajun Chen *(Zhejiang University, AZFT Joint Lab for Knowledge Engine, China)*
- **WebProtégé: A Cloud-Based Ontology Editor** 686
  Matthew Horridge, Rafael S. Gonçalves, Csongor I. Nyulas, Tania Tudorache, Mark A. Musen *(Stanford University)*
- **Content based News Recommendation via Shortest Entity Distance over Knowledge Graphs** 690
  Kevin Joseph, Hui Jiang *(York Univ)*
- **Building a Knowledge Graph for the Air Traffic Management Community** 700
  Richard M. Keller *(NASA)*
- **Scalable Knowledge Graph Construction over Text using Deep Learning based Predicate Mapping** 705
  Aman Mehta, Aashay Singhal, Kamalakar Karlapalem *(International Institute of Information Technology)*
- **How New is the (RDF) News?** 714
  Tomer Sagi, Yael Wolf *(Univ of Haifa)*, Katja Hose *(Aalborg University, Denmark)*
- **Trust and Privacy in Knowledge Graphs** 722
  Daniel Schwabe *(PUC-Rio)*
- **An Attention-based Model for Joint Extraction of Entities and Relations with Implicit Entity Features** 729
  Yan Zhou *(Univ of Chinese Academy of Sciences)*, Longtao Huang, Tao Guo *(Chinese Academy of Sciences)*, Songlin Hu *(Univ of Chinese Academy of Sciences, Chinese Academy of Sciences)*, Jizhong Han *(Chinese Academy of Sciences)*

LA-Web 2019 Chairs' Welcome

- **Misinformation Dissemination on the Web** 740
  Jussara Almeida *(UFMG)*
- **Algorithmic Bias in Rankings** 741
  Carlos Castillo *(UPF)*
- **A Link-based Approach to Detect Media Bias in News Websites** 742
  Victoria Patricia Aires, Fabiola G. Nakamura, Eduardo F. Nakamura *(UFAM)*
- **Workload Characterization of a Software-as-a-Service Web Application Implemented with a Microservices Architecture** 746
  Harold Aragon, Samuel Braganza, Edwin Boza, Jonathan Parrales, Cristina Abad *(ESPOL)*
- **A Constellation of Horrors: Analysis and Visualization of the #Cuéntalo Movement** 751
  María Soledad Bucalo, Luz Calvo, Fernando Cucchietti, David García Povedano, Artur García-Sáez, Juan Felipe Gómez, Camilo Arcadio González, Guillermo Marín *(BCS)*, Irene Meta *(BCS/IAAC)*, Patricio Reyes, Feliu Serra, Diana Fernanda Vélez *(BCS)*
- **Bitcoin Price Prediction Through Opinion Mining** 755
  Germán Cheuque Cerda, Juan L. Reutter *(Pontificia Universidad Católica de Chile Santiago, Chile)*
- **Recommender Systems for Online Video Game Platforms: the Case of STEAM** 763
  Germán Cheuque *(IMFD &Pontificia Universidad Católica Santiago, Chile)*, José Guzmán *(Pontificia Universidad Católica Santiago, Chile)*, Denis Parra *(IMFD & Pontificia Universidad Católica Santiago, Chile)*
- **The Importance of Canonical Product Attributes on User Opinions: an Empirical Investigation** 772
  Tiago de Melo *(UEA)*, Altigran S. da Silva, Edleno S. de Moura *(UFAM)*, Pável Calado *(ULisboa)*

- **Car Theft Reports: a Temporal Analysis from a Social Media Perspective** 779
  Juglar Diaz, Barbara Poblete *(UChile)*
- **Characterization of Local Attitudes Toward Immigration Using Social Media** 783
  Yerka Freire-Vidal, Eduardo Graells-Garrido *(UDD)*
- **#Default #Interactiveart #Audiencexperience** 791
  Manuela Garretón *(UC)*, Andrea Rihm *(LISCS)*
- **Global Reactions to the Cambridge Analytica Scandal: A Cross-Language Social Media Study** 799
  Felipe González *(UTFSM, Chile)*, Yihan Yu *(University of Washington, United States)*, Andrea Figueroa, Claudia López *(UTFSM, Chile)*, Cecilia Aragon *(University of Washington, United States)*
- **Implementation of the Best Practices for Data on the Web in Brazil and Costa Rica** 807
  Bernadette Farias Lóscio *(University of Pernambuco)*, Caroline Burle, Newton Calegari *(NIC.br)*
- **Mining the Relationship Between Car Theft and Places of Social Interest in Santiago Chile** 811
  Karen Oróstica *(Department of Chemical Engineering,Biotechnology and Materials, University of Chile, Chile)*, Barbara Poblete *(IMFD - Department of Computer Science University of Chile, Chile)*
- **NIFify: Towards Better Quality Entity Linking Datasets** 815
  Henry Rosales-Méndez *(DCC, University of Chile)*, Aidan Hogan, Barbara Poblete *(IMFD; DCC, University of Chile)*
- **After Brazil's General Data Protection Law: Authorization in Decentralized Web Applications** 819
  Jefferson Silva *(Pontifical Catholic University of São Paulo, Brazil)*, Newton Calegari *(Brazilian Network Information Center - NIC.br, Brazil)*, Eduardo Gomes *(Pontifical Catholic University of São Paulo, Brazil)*

The FAIR TRADE Framework for Assessing Decentralised Data Solutions

- **Towards a Decentralized, Trusted, Intelligent and Linked Public Sector: A Report from the Greek Trenches** 840
  Themis Beris *(National and Kapodistrian Univ. of Athens, Greece)*, Iosif Angelidis *(National and Kapodistrian Univ. of Athens)*, Ilias Chalkidis *(Athens Univ. of Economics and Business Athens, Greece)*, Charalampos Nikolaou *(Univ. of Oxford)*, Christos Papaloukas, Panagiotis Soursos, Manolis Koubarakis *(National and Kapodistrian Univ. of Athens)*
- **Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud** 850
  Jeremy Debattista , Judie Attard *(Trinity College, Ireland)*, Rob Brennan *(Dublin City Univ., Ireland)*, Declan O'Sullivan *(Trinity College, Ireland)*
- **Linking and Negotiating Uncertainty Theories Over Linked Data** 859
  Ahmed El Amine Djebri, Andrea G.B. Tettamanzi, Fabien Gandon *(Université Côte d'Azur, France)*
- **The FAIR TRADE Framework for Assessing Decentralised Data Solutions** 866
  John Domingue, Allan Third, Manoharan Ramachandran *(Knowledge Media Institute The Open Univ., UK)*
- **Enabling Automatic Discovery and Querying of Web APIs at Web Scale using Linked Data Standards** 883
  Franck Michel, Catherine Faron-Zucker, Olivier Corby, Fabien Gandon *(Université Côte d'Azur, France)*

LocWeb2019 Chairs' Welcome

- **Indoor Towers, DPIs, and More People in Parks at Night: New Trends in Mobile Phone Location Research** 895
  Leo Ferres *(Data Science Institute)*
- **Location Embeddings for Next Trip Recommendation** 896
  Amine Dadoun, Raphaël Troncy *(EURECOM, France)*, Olivier Ratier, Riccardo Petitti *(Amadeus SAS, France)*
- **GViz - An Interactive WebApp to Support GeoSPARQL over Integrated Building Information** 904
  Kris McGlinn, Darragh Blake, Declan O'Sullivan *(Trinity College, Ireland)*

- **Geolocation in the Browser** 913
  Thomas Steiner *(Google, Germany)*, Anssi Kostiainen *(Intel Corporation, Finland)*, Marijn Kruisselbrink *(Google, USA)*
- **Hashtag Usage in a Geographically-Local Microblogging App** 919
  Helge Reelfs, Timon Mohaupt, Oliver Hohlfeld *(RWTH Aachen University)*, Niklas Henckell *(The Jodel Venture GmbH)*

MDML'2019 Chairs' Welcome

- **Smarter Markets: Bringing Intelligence into the Exchange** 930
  David Pennock *(Microsoft Research, USA)*
- **Auction Theory from the Bidder Standpoint** 931
  Suju Rajan *(Criteo AI Lab)*, Noureddine El Karoui *(Criteo AI Lab and UC, Berkeley)*
- **Incentive-Compatible Learning of Reserve Prices for Repeated Auctions** 932
  Yash Kanoria *(Columbia Univ.)*, Hamid Nazerzadeh *(Univ. of Southern California)*
- **Learning to Bid in Revenue Maximizing Auction** 934
  Thomas Nedelec *(Criteo, ENS Paris Saclay)*, Noureddine El Karoui *(Criteo, UC, Berkeley)*, Vianney Perchet *(ENS Paris Saclay, Criteo)*

MEPDaW Chairs' Welcome

- **ML-Based Knowledge Graph Curation: Current Solutions and Challenges** 938
  Laure Berti-Equille
- **RDF Graph Stores as Convergent Datatypes** 940
  James Anderson *(Datagraph GmbH)*
- **Conflict Detection, Avoidance, and Resolution in a Non-Linear RDF Version Control System** 943
  Natanael Arndt *(AKSW, Leipzig Univ./InfAI)*, Norman Radtke *(AKSW, InfAI)*
- **Decentralized Collaborative Knowledge Management using Git** 952
  Natanael Arndt, Michael Martin *(AKSW, Leipzig Univ./InfAI)*
- **Modelling the Provenance of Linked Data Interlinks for the Library Domain** 954
  Lucy McKenna, Christophe Debruyne, Declan O'Sullivan *(ADAPT, Trinity College Dublin)*
- **TISCO: Temporal Scoping of Facts** 959
  Anisa Rula, Matteo Palmonari *(Univ. Milano-Bicocca)*, Simone Rubinacci *(Univ. Oxford)*, Axel-Cyrille Ngonga Ngomo *(Paderborn Univ.)*, Jens Lehmann *(Univ. Bonn)*, Andrea Maurino *(Univ. Milano-Bicocca)*, Diego Esteves *(Univ. Bonn)*
- **Summarizing Entity Temporal Evolution in Knowledge Graphs** 961
  Mayesha Tasnim *(RWTH Aachen)*, Diego Collarana, Damien Graux *(Fraunhofer IAIS)*, Fabrizio Orlandi *(ADAPT, Trinity College Dublin)*, Maria-Esther Vidal *(Univ. Simon Bolivar and TIB)*
- **Open traffic lights: a strategy for publishing and preserving traffic lights data** 966
  Brecht Van de Vyvere, Pieter Colpaert, Erik Mannens, Ruben Verborgh *(Ghent Univ.)*

MisInfo'19 Chairs' Welcome

- **Looking Backward to See a Way Forward** 974
  Miriam Metzger *(University of California, USA)*
- **A Topic-Agnostic Approach for Identifying Fake News Pages** 975
  Sonia Castelo *(New York Univ.)*, Thais Almeida *(Federal Univ. of Amazonas)*, Anas Elghafari, Aécio Santos, Kien Pham *(New York Univ.)*, Eduardo Nakamura *(Federal Univ. of Amazonas)*, Juliana Freire *(New York Univ.)*
- **Differences in Health News from Reliable and Unreliable Media** 981
  Sameer Dhoju, Md Main Uddin Rony *(University of Mississippi)*, Muhammad Ashad Kabir *(Charles Sturt University)*, Naeemul Hassan *(University of Mississippi)*
- **A Linked Data Model for Facts, Statements and Beliefs** 988
  Ludivine Duroyon, François Goasdoué *(Univ Rennes, Inria, CNRS, IRISA, France)*, Ioana Manolescu *(Inria and LIX (UMR 7161, CNRS and Ecole Polytechnique), France)*

- **Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking** 994
  Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, Christina Lioma *(Univ. of Copenhagen)*
- **Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-checking** 1001
  Naeemul Hassan *(Univ. of Mississippi)*, Mohammad Yousuf *(Univ. of Oklahoma)*, Md Mahfuzul Haque, Javier A. Suarez Rivas *(Univ. of Mississippi)*, Md Khadimul Islam *(Meek School of Journalism and New Media, The Univ. of Mississippi)*
- **Red Bots Do It Better:Comparative Analysis of Social Bot Partisan Behavior** 1007
  Luca Luceri *(Univ. of Applied Sciences and Arts of Southern Switzerland, and Univ. of Bern)*, Ashok Deb, Adam Badawy, Emilio Ferrara *(Univ. of Southern California)*
- **A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections.** 1013
  Caio Machado *(Univ. of Oxford)*, Beatriz Kira *(Department of Law University of Sao Paulo, Sao Paulo, Brazil)*, Vidya Narayanan, Bence Kollanyi, Philip Howard *(Univ. of Oxford)*
- **Institutional Counter-disinformation Strategies in a Networked Democracy** 1020
  Jonathan Stray *(Columbia Journalism School)*
- **Misinfosec** 1026
  Christopher R. Walker *(Marvelous AI, USA)*, Sara-Jayne Terp *(SOFWERX, USA)*, Pablo C. Breuer *(U.S. Special Operations Command, USA)*, Courtney L. Crooks, PhD *(Georgia Tech Research Institute (GTRI), USA)*

International Workshop on Modeling Social Media (MSM'2019) Chairs' Welcome

- **Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule** 1035
  Alessia Antelmi, Delfina Malandrino, Vittorio Scarano *(Universita degli Studi di Salerno, Italy)*
- **Personalized Degrees: Effects on Link Formation in Dynamic Networks from an Egocentric Perspective** 1039
  Makan Arastuie, Kevin S. Xu *(Univ. of Toledo, USA)*
- **Onto Model-based Anomalous Link Pattern Mining on Feature-Rich Social Interaction Networks** 1047
  Martin Atzmueller *(Tilburg Univ. The Netherlands)*
- **Sensing Social Media Signals for Cryptocurrency News** 1051
  Johannes Beck, Roberta Huang, David Lindner, Tian Guo, Zhang Ce, Dirk Helbing, Nino Antulov-Fantulin *(ETH Zurich, Switzerland)*
- **User Donations in a User Generated Video System** 1055
  Adele Lu Jia, Xiaoxue Shen *(China Agricultural University, China)*, Siqi Shen, Yongquan Fu, Liwen Peng *(National Univ. of Defense Technology, China)*
- **CrowdGuard: Characterization and Early Detection of Collective Content Polluters in Online Social Networks** 1063
  Ke Li, Bin Guo, Qiuyun Zhang, Jianping Yuan, Zhiwen Yu *(Northwestern Polytechnical University, China)*
- **Supervised Lexicon Extraction for Emotion Classification** 1071
  Alberto Purpura *(Univ. of Padova, Italy)*, Chiara Masiero *(Statwolf, Italy)*, Gianmaria Silvello, Gian Antonio Susto *(Univ. of Padova, Italy)*
- **Leveraging Motifs to Model the Temporal Dynamics of Diffusion Networks** 1079
  Soumajyoti Sarkar, Hamidreza Alvari, Paulo Shakarian *(Arizona State Univ.)*
- **Learning Influence Probabilities and Modelling Influence Diffusion in Twitter** 1087
  Zizhu Zhang, Weiliang Zhao, Jian Yang *(Macquarie Univ.)*, Cecile Paris, Surya Nepal *(CSIRO)*

Workshop on Subjectivity, Ambiguity and Disagreement on the Web (SAD2019) Chairs' Welcome

- **Humans, Machines and Disagreement** 1098
  Brad Klingenberg *(Stitch Fix, USA)*

- **What we talk about when we talk about crowdsourcing** 1099
  Maria Stone *(Apple, USA)*
- **Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions** 1100
  Lora Aroyo *(Google)*, Lucas Dixon, Nithum Thain, Olivia Redfield, Rachel Rosen *(Jigsaw)*
- **Discovering User Bias in Ordinal Voting Systems** 1106
  Alyssa Lees, Chris Welty *(Google)*
- **Learning to Predict Population-Level Label Distributions** 1111
  Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, Christopher Homan *(Rochester Institute of Technology, New York)*
- **TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing** 1121
  V. K. Chaithanya Manam, Dwarakanath Jampani *(Purdue Univ.)*, Mariam Zaim *(Cisco)*, Meng-Han Wu, Alexander J. Quinn *(Purdue Univ.)*
- **Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format** 1131
  Mike Schaekermann, Graeme Beaton *(Univ. of Waterloo)*, Minahz Habib, Andrew Lim *(Univ. of Toronto)*, Kate Larson, Edith Law *(Univ. of Waterloo)*
- **Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing** 1138
  Rafael Zequeira Jiménez *(TU Berlin)*, Anna Llagostera *(Rohde & Schwarz SwissQual AG)*, Babak Naderi, Sebastian Möller *(TU Berlin)*, Jens Berger *(Rohde & Schwarz SwissQual AG)*

TempWeb 2019 Chairs' Welcome

- **Large Scale Human Sensing Over Time. Challenges and Lessons Learned** 1146
  Omar Alonso *(Microsoft)*
- **Fast Interval Joins for Temporal SPARQL Queries** 1148
  Melisachew Wudage Chekol *(University Mannheim)*, Giuseppe Pirrò *(Department of Computer Science, University of Rome La Sapienza, Italy)*, Heiner Stuckenschmidt *(University of Mannheim)*
- **Epitaph or Breaking News? Analyzing and Predicting the Stability of Knowledge Base Properties** 1155
  Ioannis Dikeoulias *(Saarland University)*, Jannik Strötgen *(Bosch Center for Artificial Intelligence)*, Simon Razniewski *(Max Planck Institute for Informatics)*
- **The Power of Temporal Features for Classifying News Articles** 1159
  Lukas Lange *(Saarland University)*, Omar Alonso *(Microsoft)*, Jannik Strötgen *(Bosch Center for Artificial Intelligence)*
- **Exploring Video Game Searches on the Web** 1161
  Behrooz Mansouri *(Rochester Institute of Technology)*, Mohammad Sadegh Zahedi *(School of Electrical and Computer Engineering, University of Tehran)*, Ricardo Campos *(Polytechnic Institute of Tomar)*, Mojgan Farhoodi *(Iran Telecommunication Research Center)*

Wiki Workshop Chairs' Welcome

- **Open Personalized Navigation on the Sandbox of Wiki Pages** 1173
  Chuankai An *(Department of Computer Science, Dartmouth College, Hanover, NH, 03755)*, Daniel N. Rockmore *(Department of Computer Science and Mathematics, Dartmouth College, Hanover, NH, 03755)*
- **Understanding the Signature of Controversial Wikipedia Articles through Motifs in Editor Revision Networks** 1180
  James Ashford, Liam Turner, Roger Whitaker *(School of Computer Science and Informatics)*, Alun Preece *(Crime and Security Research Institute)*, Diane Felmlee *(Dept. of Sociology & Criminology)*, Don Towsley *(College of Information and Computer Sciences)*
- **A Graph-Structured Dataset for Wikipedia Research** 1188
  Nicolas Aspert, Volodymyr Miz, Benjamin Ricaud, Pierre Vandergheynst *(LTS2, EPFL, Station 11, Switzerland)*

- **Learning to Map Wikidata Entities To Predefined Topics** 1194
  Preeti Bhargava *(Demandbase, San Francisco, CA)*, Nemanja Spasojevic *(Youtube, San Bruno, CA)*, Sarah Ellinger *(Juul Labs, San Francisco, CA)*, Adithya Rao *(Magic Leap, Sunnyvale, CA)*, Abhinand Menon *(Beeswax, New York, NY)*, Saul Fuhrmann *(Lime Bikes, San Francisco, CA)*, Guoning Hu *(Amazon, Sunnyvale, CA)*
- **Characterizing and comparing Portuguese and English Wikipedia medicine-related articles** 1203
  Gil Domingues, Carla Teixeira Lopes *(University of Porto/INESC TEC)*
- **Thanks for Stopping By: A Study of "Thanks" Usage on Wikimedia** 1208
  Swati Goel *(Henry M. Gunn High School)*, Ashton Anderson *(University of Toronto)*, Leila Zia *(Wikimedia Foundation)*
- **Understanding Travel from Web Queries using Domain Knowledge from Wikipedia** 1212
  Chander J Iyer, Srinath Ravindran *(Yahoo! Research, Sunnyvale, California)*
- **What's in the Content of Wikipedia's Article for Deletion Discussions?** 1215
  Ali Javanmardi *(Department of Computer Science University of Western Ontario, London, ON, Canada)*, Lu Xiao *(School of Information Studies Syracuse University, Syracuse, NY, U.S.A.)*
- **Inferring Advertiser Sentiment in Online Articles using Wikipedia Footnotes** 1224
  Shaunak Mishra *(Yahoo Research)*, Aasish Pappu *(Spotify Research)*, Narayan Bhamidipati *(Yahoo Research)*
- **Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata** 1232
  Charlotte Rudnik *(LIMSI, CNRS, Orsay, France)*, Thibault Ehrhart *(EURECOM, Sophia Antipolis, France)*, Olivier Ferret *(CEA, LIST, Gif-sur-Yvette, F-91191, France)*, Denis Teyssou *(Agence France-Presse (AFP), Paris, France)*, Raphael Troncy *(EURECOM, Sophia Antipolis, France)*, Xavier Tannier *(Sorbonne Université, Inserm, LIMICS, Paris, France)*
- **Finding Prerequisite Relations using the Wikipedia Clickstream** 1240
  Mohsen Sayyadiharikandeh *(Indiana University, Bloomington)*, Jonathan Gordon *(Vassar College)*, Jose-Luis Ambite, Kristina Lerman *(USC Information Sciences Institute)*
- **How Partisanship and Perceived Political Bias Affect Wikipedia Entries of News Sources** 1248
  Khonzodakhon Umarova, Eni Mustafaraj *(Wellesley College)*
- **Detecting and Gauging Impact on Wikipedia Page Views** 1254
  Xiaoxi Chelsy Xie, Isaac Johnson, Anne Gomez *(Wikimedia Foundation)*

The Third Workshop on Women in Web Data Science (WinDS'19) Chairs' Welcome & Organization Listing

- **Information Extraction with Humans in the Loop** 1264
  Anna Lisa Gentile *(IBM Research Almaden)*
- **Responsible Data Science** 1265
  Lise Getoor *(University of California, Santa Cruz)*
- **Lightning Talk - Looking for the Movie Seven or Sven from the Movie Frozen?** 1266
  Oghenemaro Anuyah, Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera *(Boise State University)*
- **Lightning Talk - Humor Recognition in Russian Language** 1268
  Valeriya Baranova-Bolotova, Vladislav Blinov, Pavel Braslavski *(Ural Federal University)*
- **Signed Link Prediction with Sparse Data: The Role of Personality Information** 1270
  Ghazaleh Beigi *(Computer Science and Engineering, Arizona State University)*, Suhas Ranganath *(Walmart Labs)*, Huan Liu *(Computer Science and Engineering, Arizona State University)*
- **Extracting Meaningful Insights on City and Zone Levels Utilizing US Open Government Data** 1279
  Samaa Gazzaz *(King Abdulaziz University)*, Praveen Rao *(University of Missouri-Kansas)*
- **Lightning Talk - Ruuh: A Conversational Social Agent** 1286
  Meghana Joshi, Sonam Damani, Khyatti Gupta, Nitya Raviprakash *(Microsoft, India)*
- **Lightning Talk - Think Outside the Dataset: Finding Fraudulent Reviews using Cross-Dataset Analysis** 1288
  Shirin Nilizadeh *(University of Texas, at Arlington)*, Hojjat Aghakhani, Eric Gustafson, Christopher Kruegel, Giovanni Vigna *(University of California, Santa Barbara)*

- **Lightning Talk–Towards Robust Detection of Cyberbullying in Social Media** 1290
  Mengfan Yao, Charalampos Chelmis, Daphney-Stavroula Zois *(University at Albany, State University of New York)*
- **Fake News Detection: An Interdisciplinary Research** 1292
  Xinyi Zhou, Reza Zafarani *(Data Lab, EECS Department, Syracuse University)*
- **The Practice of Labeling: Everything You Always Wanted to Know About Labeling (But Were Afraid to Ask)** 1293
  Omar Alonso *(Microsoft)*
- **Crowdsourcing Inclusivity: Dealing with Diversity of Opinions, Perspectives and Ambiguity in Annotated Data** 1294
  Lora Aroyo *(Google,New York, New York, USA)*, Anca Dumitrache, Oana Inel *(Vrije Universiteit Amsterdam,Netherlands)*, Zoltán Szlávik, Benjamin Timmermans *(IBM Benelux CAS,Amsterdam, Netherlands)*, Chris Welty *(Google Research,New York, New York, USA)*
- **Designing Equitable Algorithms for the Web** 1296
  Ricardo Baeza-Yates *(NTENT & Northeastern University)*, Sharad Goel *(Stanford)*
- **Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned** 1297
  Sarah Bird *(Facebook)*, Ben Hutchinson *(Google)*, Krishnaram Kenthapadi *(LinkedIn Corp.)*, Emre Kıcıman *(Microsoft)*, Margaret Mitchell *(Google)*
- **A/B Testing at Scale: Accelerating Software Innovation** 1299
  Somit Gupta, Ronny Kohavi, Alex Deng, Jeff Omhover, Pawel Janowski *(Microsoft)*
- **Economic Theories of Distributive Justice for Fair Machine Learning** 1301
  Krishna P. Gummadi *(MPI-SWS, Saarbrücken)*, Hoda Heidari *(ETH Zürich)*
- **Tutorial on Online User Engagement: Metrics and Optimization** 1303
  Liangjie Hong *(Etsy)*, Mounia Lalmas *(Spotify)*
- **Modeling and Mining Feature-Rich Networks** 1306
  Rushed Kanawati *(LIPN, University Paris 13, Paris, France)*, Martin Atzmueller *(Tilburg University, Department of Cognitive Science & Artificial Intelligence, Tilburg, The Netherlands)*
- **Privacy-preserving Data Mining in Industry** 1308
  Krishnaram Kenthapadi *(LinkedIn Corporation)*, Ilya Mironov *(Google)*, Abhradeep Thakurta *(UC Santa Cruz)*
- **Human Mobility from theory to practice:Data, Models and Applications** 1311
  Luca Pappalardo *(ISTI-CNR)*, Gianni Barlacchi *(Univ. of Trento)*, Roberto Pellungrini *(Univ. of Pisa)*, Filippo Simini *(Univ. of Bristol)*
- **From Research Articles to Knowledge Graphs** 1313
  Vayianos Pertsas, Panos Constantopoulos *(Athens Univ. of Economics and Business)*
- **Tutorial: Sequence-Aware Recommender Systems** 1316
  Massimo Quadrana *(Pandora Media, Oakland, USA)*, Dietmar Jannach *(Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria)*, Paolo Cremonesi *(Politecnico di Milano, Milan, Italy)*
- **Scalable Subgraph Counting: The Methods Behind The Madness** 1317
  Comandur Seshadhri *(UC Santa Cruz)*, Srikanta Tirthapura *(Iowa State Univ.)*
- **Concept to Code: Deep Learning for Fashion Recommendation** 1319
  Omprakash Sonie *(Flikart)*, Muthusamy Chelliah *(Flipkart)*, Shamik Sural *(IIT Kharagpur)*
- **Representation Learning on Networks: Theories, Algorithms, and Applications** 1321
  Jie Tang *(Tsinghua Univ.)*, Yuxiao Dong *(Microsoft)*
- **Continuous Analytics of Web Streams** 1323
  Riccardo Tommasini *(Politecnico di Milano, Italy)*, Robin Keskisärkkä *(Linköping University, Sweden)*, Jean-Paul Calbimonte *(Haute Ecole Spécialisée de Suisse Occidentale (HES-SO), Switzerland)*, Eva Blomqvist *(Linköping University, Sweden)*, Emanuele Della Valle *(Politecnico di Milano, Italy)*, Albert Bifet *(LTCI, Telecom ParisTech, France)*
- **Socially Responsible Natural Language Processing** 1326
  Yulia Tsvetkov *(Carnegie Mellon Univ.)*, Vinodkumar Prabhakaran *(Google LLC)*, Rob Voigt *(Stanford Univ.)*
- **The Challenge of API Management: API Strategies for Decentralized API Landscapes** 1327
  Erik Wilde *(Good API)*, Mike Amundsen *(Amundsen Training)*

- **Deep Chit-Chat: Deep Learning for Chatbots** 1329
  Wei Wu *(Microsoft Corporation, China)*, Rui Yan *(Institute of Computer Science and Technology, Peking University, China)*
- **WWW'19 Tutorial on Explainable Recommendation and Search** 1330
  Yongfeng Zhang *(Rutgers Univ.)*, Jiaxin Mao *(Tsinghua Univ.)*, Qingyao Ai *(University of Massachusetts)*

Author Index 1332

Conference Organization

Conference Chairs:	Ricardo Baeza-Yates (NTENT & Northeastern University, USA) Leila Zia (Wikimedia Foundation, USA)
Program Chairs:	Ling Liu (Georgia Tech, USA) Ryen White (Microsoft Research, USA)
Local Organizing Chairs:	Chris Brooks (University of San Francisco, USA) Regan Robertson (Executivevents, USA) Bebo White (SLAC National Accelerator Laboratory, USA)
Treasurer:	Fabian Suchanek (Télécom ParisTech University, France)
Tutorial Chair:	Ashish Goel (Stanford University, USA)
Workshop Chairs:	Sihem Amer-Yahia (CNRS & Université Grenoble Alpes, France) Mohammad Mahdian (Google, USA)
Demo Chairs:	Amin Mantrach (Criteo, USA) Fabrizio Silvestri (Facebook, UK)
PhD Symposium Chairs:	Geert-Jan Houben (TU Delft, The Netherlands) Kristina Lerman (Information Sciences Institute, USA)
BIG Chairs:	Vanja Josifovski (Airbnb, USA) Liane Lewin-Eytan (Amazon, Israel)
Web of Health Chairs:	Evgeniy Gabrilovich (Google, USA) Nigam H. Shah (Stanford University, USA)
Future of the Open Web Chairs:	Stephen LaPorte (Wikimedia Foundation, USA) Robert West (EPFL, Switzerland)
Designing an Ethical Web Chairs:	Sharad Goel (Stanford University, USA) Deirdre Mulligan (UC Berkeley, USA)
Sponsorship Chairs:	Nuria Oliver (Vodafone Research and DataPop Alliance, Spain) Jie Tang (Tsinghua University, China) Kuansan Wang (Microsoft, USA)
Student Relations Chair:	Ben Steichen (Cal Poly Pomona, USA)

Student Scholarship Chairs: Lora Aroyo (VU Amsterdam, The Netherlands)
Diego Sáez-Trumper (Wikimedia Foundation, Spain)

Communication Chairs: Dawn Anderson (Move It Marketing, UK)
Brent Hecht (Northwestern University, USA)
Xin Fu (Facebook, USA)

Proceedings Chair: Julian McAuley (UC San Diego, USA)

Poster Chairs: Ashton Anderson (University of Toronto, Canada)
Joe DeBlasio (Google, USA)
Reza Zafarani (Syracuse University, USA)

Track Chairs: Tarek Abdelzaher (University of Illinois Urbana-Champaign, USA)
Karl Aberer (École Polytechnique Fédérale de Lausanne, Switzerland)
Omar Alonso (Microsoft, USA)
Elisa Bertino (Purdue University, USA)
Paolo Boldi (University of Milan, Italy)
Carlos Castillo (Pompeu Fabra University, Spain)
Munmun De Choudhury (Georgia Institute of Technology, USA)
Ingemar Cox (University College London, UK)
Nick Craswell (Microsoft, USA)
Debora Donato (Mix, USA)
Aristides Gionis (Aalto University, Finland)
Arun Iyengar (IBM, USA)
Emre Kiciman (Microsoft, USA)
Ashish Kundu (IBM Research, USA)
Edith Law (University of Waterloo, Canada)
Hang Li (Bytedance AI Lab, Hong Kong)
Tie-Yan Liu (Microsoft, China)
Bing Liu (University of Illinois Chicago, USA)
Yoelle Marek (Amazon Research, Israel)
Peter Mika (Independent Consultant, Spain)
Alice Oh (Korea Advanced Institute of Science and Technology, South Korea)
Michael Rabinovich (Case Western Reserve University, USA)
Filip Radlinski (Google, UK)
Pavel Serdyukov (Yandex, Russia)
Sergei Vassilvitskii (Google, USA)
Nalini Venkatasubramanian (University of California Irvine, USA)
Cong Yu (Google, USA)

THE WEB CONFERENCE

Incorporating System-Level Objectives into Recommender Systems

Himan Abdollahpouri
Department of Information Science
University of Colorado Boulder
himan.abdollahpouri@colorado.edu

ABSTRACT

One of the most essential parts of any recommender system is personalization– how acceptable the recommendations are from the user's perspective. However, in many real-world applications, there are other stakeholders whose needs and interests should be taken into account. In this work, we define the problem of multi-stakeholder recommendation and we focus on finding algorithms for a special case where the recommender system itself is also a stakeholder. In addition, we will explore the idea of incremental incorporation of system-level objectives into recommender systems over time to tackle the existing problems in the optimization techniques which only look for optimizing the individual users' lists.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*.

KEYWORDS

Recommender Systems; Multistakeholder recommendation; Multi-objective recommendation

ACM Reference Format:
Himan Abdollahpouri. 2019. Incorporating System-Level Objectives into Recommender Systems. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314201

1 PROBLEM

Recommender systems provide personalized information access, supporting e-commerce, social media, news, and other applications where the volume of content would otherwise be overwhelming. They have become indispensable features of the Internet age, found in systems of many kinds. Even search engines, the most fundamental web applications, have become increasingly personalized in their provision of results, to the extent that they can also be considered recommender systems.

One of the defining characteristics of recommender systems is personalization. Recommender systems are typically evaluated on their ability to provide items that satisfy the needs and interests of the end user. Such focus is entirely appropriate–Users would not make use of recommender systems if they believed such systems were not providing items that matched their interests. Still, it is also clear that, in many recommendation domains, the user for whom recommendations are generated is not the only stakeholder in the recommendation outcome. Other users, the providers of products, and even the system's own objectives may need to be considered. Fairness and balance are important examples of system-level objectives, and these social-welfare oriented goals may at times run counter to individual preferences. Sole focus on the end user hampers developers' ability to incorporate such objectives into recommendation algorithms and system designs.

The integration of the perspectives of multiple parties into recommendation generation and evaluation is the underlying goal of the new sub-field of *multistakeholder recommendation* [3, 6, 7, 11]. The goal of a recommender system in a multistakeholder environment is, therefore, to generate recommendations taking all the stakeholder's needs and preferences into account.

1.1 Multistakeholder Recommendation

In many recommender systems, there could exist a variety of different stakeholders. However, we can usually see three main stakeholders:

- **Consumers *C*:** The consumers are those who receive the recommendations. They are the individuals whose choice or search problems bring them to the platform, and who expect recommendations to satisfy those needs.
- **Providers *P*:** The providers are those entities that supply or otherwise stand behind the recommended objects and possibly gain utility from the consumer's choice.
- **System *S*:** The final category is the platform itself, which has created the recommender system in order to match consumers with providers and has some means of gaining benefit from successfully doing so. The platform may be a retailer, e-commerce site, broker, or other venue where consumers seek recommendations.

Typically, the system will attempt to satisfy the needs of at least one type of stakeholder by offering recommendations tailored (at least in part) to their preferences. When the stakeholder is the consumer, this is the most common personalized recommendation scenario. When there is personalization for both the consumer and the provider, we have a reciprocal recommendation.

In many real-world contexts, the system may gain some utility when recommending items, which could be in the form of a simple aggregate of the other stakeholders' utilities. In many e-commerce settings, the system will get a commission for each sale, and such benefits can be considered together with personalization [11].

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314201

Alternatively, the system may seek to tailor outcomes specifically to achieve particular objectives. For example, an educational site may view the recommendation of learning activities as a curricular decision and seek to have its recommendations fit a model of student growth and development. Its utility may, therefore, be more complex than a simple aggregation of those of the other stakeholders. One way to think about a multistakeholder recommendation is to look at it as a two-stakeholder system: the user and the system. In other words, in many applications, it is possible to aggregate the utilities of other stakeholders within the system utility and therefore optimizing for the system utility would, indirectly, optimize for their utility as well. Therefore, in this work, we focus on a scenario where we have two main stakeholders, the end-user and the system.

In contrast to a single-stakeholder perspective which has been the case, for the most part, in recommender systems research, in this work we are interested in finding algorithms and mechanisms to incorporate system-level objectives into the recommendations.

That being said, there are several challenges that need to be investigated:

- How can the system-level preferences be incorporated into the recommendation generation?
- What are the different types of system-level preferences and how does the solution for incorporating each of them differs from the others?
- What is the right balance or trade-off between user preferences and system preferences? Does this balance vary depending on the domain?
- How can the system learn from the optimization in previous steps to do a better job at a particular time?

2 STATE OF THE ART

There is a large body of recent work on incorporating diversity, novelty, long tail promotion and other metrics as additional objectives for recommendation generation and evaluation. See, for example, [2, 15]. There is also a growing body of work on combining multiple objectives using constraint optimization techniques, including linear programming. See, for example, [8, 9]. These techniques provide a way to limit the expected loss on one metric (typically accuracy) while optimizing for another, such as diversity. The complexity of these approaches increases exponentially as more constraints are considered, making them a poor fit for the general multistakeholder case. Also, for the most part, multiobjective recommendation research concentrates on maximizing multiple objectives for a single stakeholder, the end user.

3 PROPOSED APPROACH

In this work, we will focus on one specific case of multistakeholder recommendation which is considering users and the system as two different stakeholders. That being said, we realized there is a substantial amount of work on using different types of non-accuracy metrics in recommender systems each of which used a different method to do the recommendation generation. For example, most of the work on recommendation list diversification, price optimization, fairness aware recommendation and so on use different algorithms to achieve the desired set of recommendation based on what evaluation metrics they use. In other words, there is a lack of understanding of what all these different problems have in common or from what aspects they differ. What is missing is the need for a general way of looking at extra objectives (i.e. system objectives) in recommender systems and to define all different types of such objectives and develop algorithms which can be used in problems with a similar type of objectives. What we propose is to have a standard framework that could be the home for all these different problems. To start, we need to categorize different types of system-level objectives. One main categorization in terms of the type of optimization that could be used to solve such problems is based on the scope of the system objective. Therefore, we divide every system-level objective into two main groups:

Local objectives: in this group we have objectives that could be optimized within each user's recommendation list. In other words, there is no need to look at the entire user population in order to do the optimization. List diversification and novelty (from the user's perspective) belongs to this group.

Global objectives: The second group is the type of objectives that need to be taken care of not only based on the recommendations within each user's list but also the recommendation lists of other users as well. Objectives like long-tail promotion and fair exposure belong to this category.

In addition to this two classes of objectives, we can also define several other types of objectives based on whether the system's focus is on 1) items, 2) users or maybe 3) both at the same time. We define different types of system objectives with respect to which entity the system is focusing on as follow:

User-item related: The system could have a certain goal regarding which user should get what recommendation. This is more like a parental control which means, in addition to what users like, the system also has to decide whether that item suits the user or not. An example would be in educational recommendation where the system might not want to recommend a program to a user even though she likes it or, recommend a certain program even though she may not like it the most compared to other programs.

user related: When the system wants the users who receive a certain item (items) as the recommendation to have certain properties. For example, in an educational recommender system, the goal might be to make sure the users who receive the recommendation are %50 male and %50 female. Many of the fairness-aware recommendation problems fall in this category.

item related: This is when the system has a certain goal regarding a certain group of items. For example, long-tail recommendation, which is a well-known problem, is when the system wants to promote long-tail items more often.

4 RESULTS AND THE PROGRESS SO FAR

As a very specific case of system-level objective, we have worked on the problem of long-tail promotion (i.e. controlling popularity bias) in recommender systems. Popularity bias refers to the situation in which there is an over-concentration of the recommendations of certain items, the popular ones. Controlling popularity bias is important for many recommender systems as it affects the fairness

of the platform and it also improves the novelty of the recommendations. For more information on popularity bias see [1, 2]main. We wanted to go a bit deep in this problem and find different types of solutions for it to make a better sense of this type of objective and to be able to generalize for similar problems. Our aim is, however, to find a general class of solutions for other types of system objectives for the next steps of this research.

In order to control the effect of popularity bias in recommender systems we have used two approaches: One way is to modify the underlying recommendation algorithm such that we take the popularity of items into account in addition to the relevance of those items for generating the predicted ratings. Therefore, the final list is both personalized and also controlled for popularity bias.

Another approach is to keep the existing algorithm untouched but add an extra level of item filtering on the predicted ratings. In this method, the algorithm tries to re-rank the items based on both their accuracy (predicted rating) and popularity.

4.1 Technique 1: Model based Approach: Fairness-aware Regularization

In this approach, we have explored the use of regularization to control the popularity bias of a recommender system. We start with an optimization objective of the form:

$$\min_{P,Q} acc(P,Q) + \lambda reg(P,Q) \tag{1}$$

where $acc(.)$ is the accuracy objective, $reg(.)$ is a regularization term, and λ is a coefficient for controlling the effect of regularizer.

Our goal, therefore, is to identify a regularization component of the objective function that will be minimized when the distribution of recommendations is fair in terms of popular and non-popular items.

We define two sets of items (Γ') and (Γ) corresponding to the popular and non-popular items, and define a co-membership matrix D, over these sets. For any pair of items i and j, $d(i,j) = 1$ if i and j are in the same set and 0 otherwise. We define intra-list binary unfairness (ILBU) as the average value of $d(i,j)$ across all pairs of items i, j.

$$ILBU(L_u) = \frac{1}{N(N-1)} \sum_{i,j \in L_u} d(i,j) \tag{2}$$

where N is the number of items in the recommendation list. The fairest list is one that contains equal numbers of items from each set, which can be easily seen.

For more details on how the optimization is solved please refer to [2].

4.2 Technique 2: Post processing Re-ranking

4.2.1 xQuAD. For the second approach, we used a post-processing re-ranking technique to diversify the recommendation lists in terms of popular and unpopular items. We have used EXplicit Query Aspect Diversification (xQuAD) [12] which explicitly accounts for the various aspects associated with an under-specified query. In the context of search engines, items are selected iteratively by estimating how well a given document satisfies an uncovered aspect. We build on the *xQuAD* model to control popularity bias in recommendation algorithms. We assume that for a given user u, a ranked recommendation list R has already been generated by a base recommendation algorithm. The task of the modified xQuAD method is to produce a new re-ranked list S ($|S| < |R|$) that manages popularity bias while still being accurate.

The new list is built iteratively according to the following criterion:

$$P(v|u) + \lambda P(v, S'|u) \tag{3}$$

where $P(v|u)$ is the likelihood of user $u \in U$ being interested in item $v \in V$, independent of the items on the list so far as predicted by the base recommender. The second term $P(v, S'|u)$ denotes the likelihood of user u being interested in an item v as an item not in the currently generated list S.

Intuitively, the first term incorporates accuracy while the second term promotes diversity between two different categories of items (i.e. short head and long tail). The parameter λ controls how strongly controlling popularity bias is weighted in general. The item that scores most highly under the equation 3 is added to the output list S and the process is repeated until S has achieved the desired length.

For more details on these two techniques refer to [2] and [4].

we tested our proposed algorithms on two public datasets. We show the result for one dataset due to lack of space but a, more or less, a similar result could be seen for the other dataset. The dataset for which you can see the result in figure 1, is the Epinions dataset which is gathered from a consumers opinion site where users can review items [10]. This dataset has the total number of 664,824 ratings given by 40,163 users to 139,736 items. We split the items in both datasets into two categories: long-tail (Γ) and short head (Γ')in a way that short head items correspond to %80 of the ratings while long-tail items have the rest of the %20 of the ratings.

4.3 Evaluation

The experiments compare four algorithms. Since we are concerned with ranking performance, we chose as our baseline algorithm RankALS, a pair-wise learning-to-rank algorithm. We used the output from RankALS as input for the re-ranking technique described above. Note that, we developed two versions of the re-ranking algorithm: Binary xQuAD and Smooth xQuAD, marked *Binary* and *Smooth* in the figures. The difference between the two as you can read in more details in [4] is in how they check for whether an item category (popular or non-popular) is already covered by the list or not. The binary version only returns 0 (not covered) or 1 (covered) while the smooth version returns what percentage of the list is covered with each item category. Also, the model-based approach is shown by *LT_Reg* in the figure. We compute lists of length 100 from RankALS and pass these to the re-ranking algorithms to compute the final list of 10 recommendations for each user. In order to evaluate the effectiveness of algorithms in mitigating popularity bias, we used four different metrics each of which measures a certain aspect of long-tail promotion.

Average Recommendation Popularity (ARP): This measure from [14] calculates the average popularity of the recommended items in each list. For any given recommended item in the list, we measure the average number of ratings for those items. More

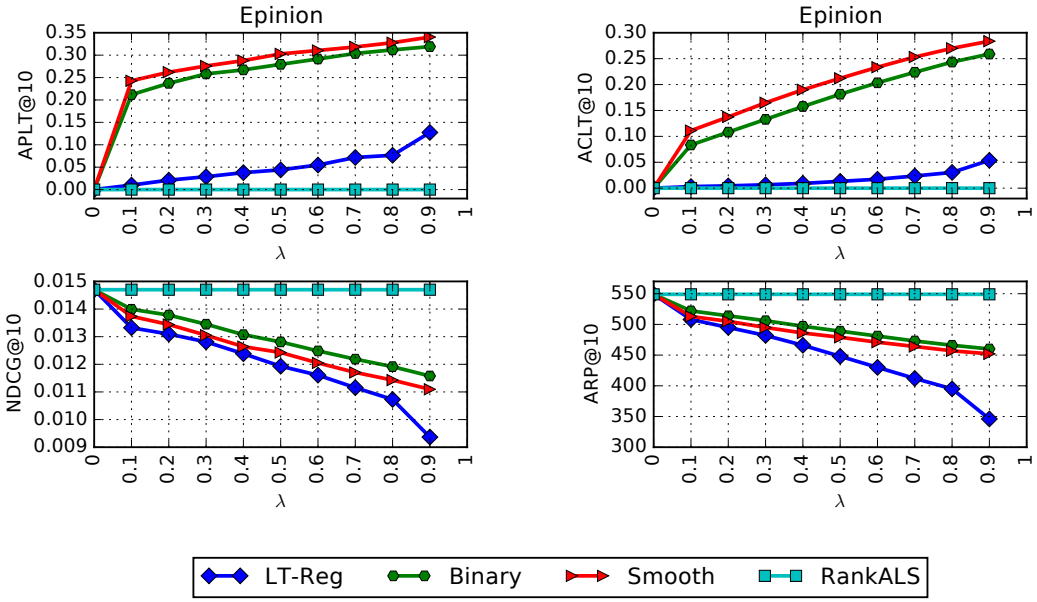


Figure 1: Results for the Epinions dataset

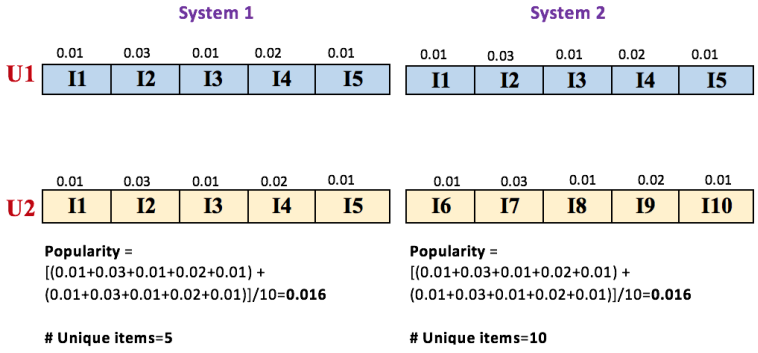


Figure 2: The recommendation lists of length 5 to different users U1 and U2 generated by two different recommender systems. Both systems have similar sum of popularity, but system 2 has covered 10 items while system 1 has only covered 5.

formally:

$$ARP = \frac{1}{|U_t|} \sum_{u \in U_t} \frac{\sum_{i \in L_u} \phi(i)}{|L_u|} \tag{4}$$

where $\phi(i)$ is the number of times item i has been rated in the training set. L_u is the recommended list of items for user u and $|U_t|$ is the number of users in the test set.

Average Percentage of Long Tail Items (APLT): As used in [2], this metric measures the average percentage of long tail items in the recommended lists and it is defined as follows:

$$APLT = \frac{1}{|U_t|} \sum_{u \in U_t} \frac{|\{i, i \in (L_u \cap \Gamma)\}|}{|L_u|} \tag{5}$$

This measure gives us the average percentage of items in users' recommendation lists that belong to the long tail set.

Average Coverage of Long Tail items (ACLT): This is the metric we introduced in [4] which measures how much exposure long-tail items get in the entire recommendations. One problem with *APLT* is that it could be high even if all users get the same set of long tail items. *ACLT* measures what fraction of the long-tail items the recommender has covered:

$$ACLT = \frac{1}{|U_t|} \sum_{u \in U_t} \sum_{i \in L_u} \mathbb{1}(i \in \Gamma) \tag{6}$$

where $\mathbb{1}(i \in \Gamma)$ is an indicator function and it equals to 1 when i is in Γ. This function is related to the *Aggregate Diversity* metric of [5] but it looks only at the long-tail part of the item catalog.

In addition to the aforementioned long tail diversity metrics, we also evaluate the accuracy of the ranking algorithms in order to examine the diversity-accuracy trade-offs. For this purpose we

use the standard *Normalized Discounted cumulative Gain* (*NDCG*) measure of ranking accuracy.

Figure 2 shows the results for the Epinions dataset across the different algorithms using a range of values for λ. (Note that the LT-Reg algorithm uses the parameter λ to control the weight placed on the long-tail regularization term.) All results are averages from five-fold cross-validation using a %80 -%20 split for train and test, respectively. As expected, the diversity scores improve for all algorithms, with some loss of ranking accuracy. Differences between the algorithms are evident, however. The exposure metric (ACLT) plot shows that the two re-ranking algorithms, and especially the Smooth version, are doing a much better job of exposing items across the long-tail inventory than the regularization method. The ranking accuracy shows that, as expected, the Binary version does slightly better as it performs minimal adjustment to the ranked lists. LT-Reg is not as effective at promoting long-tail items, either by the list-wise APLT measure or by the catalog-wise ACLT.

5 CONCLUSIONS AND FUTURE WORK

Working on tackling popularity bias as an example of system-level objective, we learned that the system objective could be incorporated in three different ways which we only explored two of them: 1) Model-based multi-objective optimization, 2) re-ranking approach, and 3) data manipulation which we have not yet worked on. The work we have done so far, might not be, just by itself, a great contribution to the field as there are similar approaches for tackling popularity bias but it inspired us to look at the problem in a more general way as popularity bias is not the only objective that a system could have. For example, *user-level list diversity*, *aggregate diversity*, *profit maximization*, *fairness-aware recommendation*, to name just a few, all could be considered as system-level objectives which should be incorporated into the recommendation process. So one future work is to find a general framework by which we can incorporate any type of system-level objectives into the recommender systems.

Another important observation we made which is also a possibility for future work is that almost all the works on long-tail promotion have done the optimization on individual's recommendation lists hoping this will lead to an overall optimization within the entire user base. However, as we saw in our experiments (ACLT vs ARP), a system could have a low overall sum of popularity for the recommendations but might not have done a good job of promoting enough long-tail items. The reason is optimizing the individual's recommendation list (local optimization) does not guaranty an overall optimization (global optimization) and, therefore, a more dynamic and temporal optimization technique is needed. Figure 2 shows a simple scenario where two systems have generated recommendations for two users. The popularity values for each item is shown on the top of each recommended item. A popularity value of 0.01 means the item has been rated by %1 of the users. As you can see, both systems have the same sum of popularity but one has covered 10 items while the other has only covered 5. This example emphasizes the need for better incorporation of system-level objectives into recommendations. One possibility to address this problem is to use incremental optimization [13] over time to balance system and user-level objectives. Figure 3 shows a schematic view of how the optimization could be done over time such that, at any given time T, the recommender system optimizes the objective function by also looking at what the outcome for optimization was at previous times.

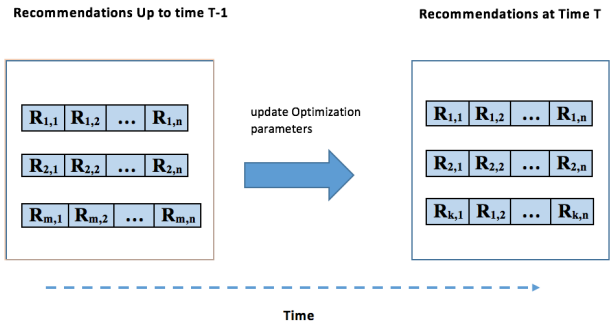


Figure 3: Dynamic optimization of the recommendations over time

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity Bias in Ranking and Recommendation. In *In AAAI/ACM Conference on AI, Ethics, and Society (AIES'19) January 27–28, 2019, Honolulu, HI, USA.* ACM.

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning to Rank Recommendation. In *Proceedings of the 11th ACM conference on Recommender systems.*

[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Recommender Systems as Multistakeholder Environments. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP'17).* Bratislava, Slovakia, To appear.

[4] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-ranking.. In *FLAIRS Conference.* To appear.

[5] G. Adomavicius and Y.O. Kwon. 2012. Improving aggregate recommendation diversity using ranking-based techniques. *Knowledge and Data Engineering, IEEE Transactions on* 24, 5 (2012), 896–911.

[6] Robin Burke and Himan Abdollahpouri. 2016. Educational Recommendation with Multiple Stakeholders. In *Third International Workshop on Educational Recommender Systems.* Omaha.

[7] Robin Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. 2016. Towards Multi-Stakeholder Utility Evaluation of Recommender Systems. In *Proceedings of the International Workshop on Surprise, Opposition, and Obstruction in Adaptive and Personalized Systems (SOAP 2016).* ACM.

[8] Tamas Jambor and Jun Wang. 2010. Optimizing Multiple Objectives in Collaborative Filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10).* ACM, New York, NY, USA, 55–62.

[9] Yuanchun Jiang and Yezheng Liu. 2012. Optimization of online promotion: a profit-maximizing model integrating price discount and product recommendation. *International Journal of Information Technology & Decision Making* 11, 05 (2012), 961–982.

[10] Paolo Massa and Paolo Avesani. 2007. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems.* ACM, 17–24.

[11] Phong Nguyen, John Dines, and Jan Krasnodebski. 2017. A Multi-Objective Learning to re-Rank Approach to Optimize Online Marketplaces for Multiple Stakeholders. *arXiv preprint arXiv:1708.00651* (2017).

[12] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web.* ACM, 881–890.

[13] Alexa Megan Sharp. 2007. *Incremental algorithms: solving problems in a changing world.* Cornell University.

[14] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *Proceedings of the VLDB Endowment* 5, 9 (2012), 896–907.

[15] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web.* ACM, 22–32.

Identification of the Source(s) of Misinformation Propagation Utilizing Identifying Codes

Kaustav Basu
NetXT Lab, SCIDSE, Arizona State University
kaustav.basu@asu.edu

ABSTRACT

With billions of users, social networks have become the go to platform for information diffusion for news media outlets. Lately, certain entities (users and/or organizations) have been active in generating misinformation in order to attract users to their respective websites, to generate online advertisement revenues, to increase followers, to create political instability, etc. With the increasing presence of misinformation on social networks, it is becoming increasingly difficult to not only distinguish between information and misinformation, but also, to identify the source(s) of misinformation propagation. This effort reviews my doctoral research on identifying the source(s) of misinformation propagation. Particularly, I utilize the mathematical concept of Identifying Codes to *uniquely* identify users who become active in propagating misinformation. In this paper, I formally present the computation of the Minimum Identifying Code Set (MICS) as a novel variation of the traditional Graph Coloring problem. Furthermore, I present an Integer Linear Program for the computation of the MICS. I apply the technique on various anonymous Facebook network datasets and show the effectiveness of the approach.

KEYWORDS

Identifying Codes; Misinformation Propagation; Unique Identification; Graph Analytics

ACM Reference Format:
Kaustav Basu. 2019. Identification of the Source(s) of Misinformation Propagation Utilizing Identifying Codes. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314200

1 PROBLEM

The emergence of social networking has revolutionized the communication domain. It has successfully overcome the geographical barrier and has connected users with their friends and family members, located on other the side of the planet. With the advantages of a social network being in abundance, it also suffers from a few major drawbacks. In recent times, the inability of a social networking company to maintain the privacy of a user has become front page news. Moreover, adversaries are trying to utilize the connectivity and reach of a social network, to spread misinformation. The identification of the source of misinformation propagation is a chief problem in today's social network analysis. Hence, as a part of this research effort, I address the problem of *unique* identification of the source(s) of misinformation propagation, by utilizing the mathematical concept of *Identifying Codes*.

With the evolution of networks over the past decade, social networks have been utilized by news outlets and media houses for the propagation of a respective outlet's/house's agenda. The greater the utilization, the greater the reliance of the user on the social media content. Certain entities (individuals or organizations) on social networking websites have taken advantage of this reliance, to propagate misinformation. Agencies (or monitors) such as Politifact and Media Bias/Fact Check (MBFC), are trusted and verified agencies who can determine if a news article is informed or misinformed. Even with the existence of such monitors, it is becoming increasingly difficult to distinguish between informed and misinformed articles, due to the sheer volume of such articles on social networking websites. Moreover, such fact checking agencies do not verify news articles in real time. As a result, various research groups have undertaken the study of misinformation propagation to primarily (i) develop algorithms to detect whether a news article is informed or misinformed, and (ii) identify the source(s) of misinformation propagation.

Mathematically, a social network can be expressed as a graph $G = (V, E)$, which consists of a set of vertices V and a set of edges E, where two vertices $u, v \in V$ are said to be connected if there is an edge $e \in E$ between them. The set of vertices can be thought of as entities, such as individuals, organizations, locations, etc. and the set of edges can be thought of relationships between these entities. For example, in the case of a social network, users of the social network form the set of vertices, and friendship between two users form the set of edges.

As a part of my doctoral research, I am currently studying the accurate identification of source(s) of misinformation propagation. I am in the process of utilizing the mathematical concept of Identifying Codes in order to accurately identify users/organizations on the social network, by placing the minimum number of such monitors or sensors (fact checking agencies such as Politifact or MBFC) on a social network. It may be noted that in this effort, I not only study the propagation of misinformation in the network, but also present a novel approach which can accurately identify the source(s) of misinformation propagation.

One of the most frequently studied problems in this context is the Sensor Placement Optimization problem. It may be noted that a sensor in this context refers to the monitoring agencies, such as Politifact and MBFC. The goal of the sensor optimization problem is to find the smallest set of nodes on which sensors must be

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314200

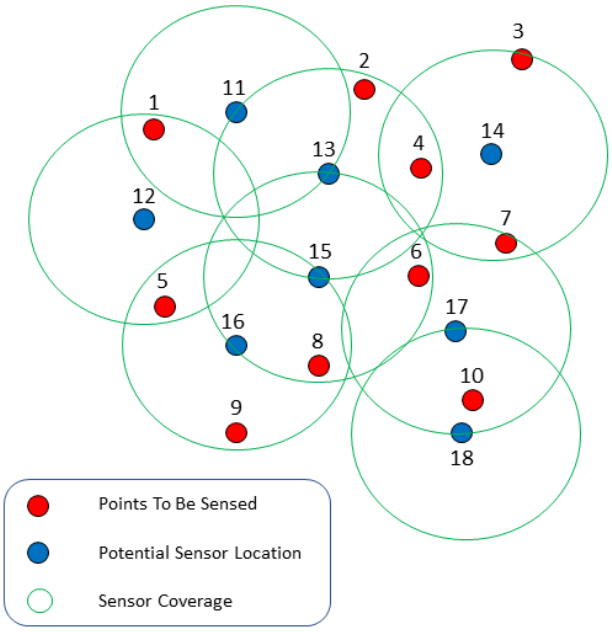


Figure 1: Potential Sensors and Sensing Locations

placed, so that all the nodes in the network can be sensed. In other words, every node should be under the coverage area of at least one deployed sensor. In the context of misinformation propagation in a social network, find the smallest set of users who should be monitored by the fact checking entities entities, so that all the users in the social network are under the coverage of at least one monitored user.

Although a number of studies on sensor placement optimization problem follow the set cover formulation to find a solution, the set cover based approach has a serious limitation on the accurate identification of the node, where some abnormality is detected by one or more of the deployed sensors. We illustrate this point with the help of an example. In Fig. 1, the ten red points (numbered from 1-10) indicate the points to be sensed (monitored), the eight blue points (numbered from 11-18) indicate the potential locations where the sensors can be deployed and the green circles (centered on each blue point) indicate the coverage area of a sensor deployed at that blue point. We can easily construct a bipartite graph $G = (V_1 \cup V_2, E$, from the set of points in Fig. 1. The set V_1 could denote the set of red points and the set V_2 could denote the set of blue points. If a red point is within the sensing range of a blue point, then we have an edge between the two points. The resulting graph from Fig. 1 is illustrated in Fig. 2. In this example, it can be verified that if the set cover approach was followed, and sensors were deployed in locations 11, 13, 14, 16 and 17, then all points from 1 to 10 will be within the sensing range of at least one sensor. Specifically, the points (1-10) covered (sensed) by the sensors 11, 13, 14, 16, 17 are shown in Table 1. In Table 2, we present the sensors that are actually sensing the points 1-10, using the Set Cover approach.

The serious limitation of the set cover based approach to optimal sensor placement problem is that, it may fail to uniquely identify

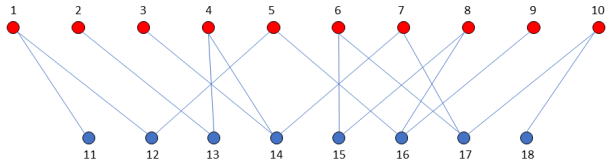


Figure 2: Network Corresponding To Fig. 1

Table 1: Points Covered By Each Sensor

Sensor Location	Points Sensed	Sensor Location	Points Sensed
11	1	15	6, 8
12	1, 5	16	5, 8, 9
13	2,4	17	6, 7, 10
14	3,4,7	18	10

Table 2: Sensors Covering Each Point

Points Sensed	Sensor Location	Points Sensed	Sensor Location
1	11	6	17*
2	13	7	14, 17
3	14	8	16*
4	13, 14	18	16*
5	16**	10	17*

Table 3: Sensors Covering Each Point

Points Sensed	Sensor Location	Points Sensed	Sensor Location
1	12	6	15, 17
2	13	7	14, 17
3	14	8	15, 16
4	13, 14	18	16
5	12, 16	10	17

the point where an abnormality is detected by the sensor. We elaborate this point with the results shown in Tables 1 and 2. In this example, sensors were deployed at locations 11, 13, 14, 16, 17 and this deployment ensured that all points to be sensed were within the coverage area of at least one sensor. Suppose a control center has five indicator lamps A, B, C, D and E corresponding to five sensors located at 11, 13, 14, 16 and 17. If the sensor does not sense an abnormality at the point it is sensing, then the corresponding lamp is lit green. If a sensor senses an abnormality at a point, then the corresponding lamp turns red. From Table 2, it can be seen that points 6 and 10 are sensed by sensor 17 only, and points 5, 8 and 9 are sensed by sensor 16 only. The implication of this is that if lamp E (corresponding to sensor 17) turns red, then it will not be possible to ascertain if the abnormality was detected at point 6 or 10. Similarly, if lamp D (corresponding to sensor 16) turns red, then it will not be possible to ascertain if the abnormality was detected at point 5 or 8 or 9.

This limitation of failure to uniquely identify the point where abnormality is detected by the sensor, can be overcome by deployment of additional sensors. In this example, instead of deploying sensors at locations 11, 13, 14, 16 and 17, if they were deployed at locations 12, 13, 14, 15, 16 and 17, then each point would have been sensed in the way as shown in Table 3. It may be noted that deployment of six sensors, instead of five, avoids the problem of failure of unique identification of points where abnormality is detected. The mathematical foundation of computing the least number of sensors needed to uniquely identify locations (or nodes) where abnormality is detected, is called *Identifying Codes*.

2 RELATED WORK

Over the past half decade, there has been significant studies in the field of misinformation propagation. In this section, I highlight few efforts which motivated my research. Abbasi in [1], studied a method to measure user credibility in social media based on the user's profile. Ciampaglia in [5], showed that the complexities of human fact checking can be approximated by finding the shortest path between concept nodes under properly defined semantic proximity metrics on knowledge graphs. The problem of fake news mitigation was mapped to the reinforcement learning framework, with the goal of optimizing the actions for maximal total reward under budget constraints in [6]. Shi in [13], viewed link-prediction task in a knowledge graph to accurately determine the veracity of a fact. Shu *et. al.* in [14] presented a survey of detecting fake news on social media. Real-world datasets measuring users trust level on fake news was constructed in [15]. Tachhini *et. al.* in [17] showed that Facebook posts can be classified with high accuracy as hoaxes or non-hoaxes on the basis of the users who "liked" them.

Karpovsky *et. al.* introduced the concept of Identifying Codes in [7] and provided results for Identifying Codes for graphs with specific topologies, such as binary cubes and trees. Using Identifying Codes, Laifenfeld *et. al.* studied covering problems in [8]. Charon *et. al.* studied studied complexity issues and showed that in several types of graph, the problem is NP-hard in [4]. Approximation algorithms for computation of Identifying Codes for some special types of graphs are presented in [18] and [16]. Ray *et. al.* studied location detection problem in emergency sensor networks, using Identifying Codes [10]. In this paper, they also introduced the concept of robust Identifying Codes to deal with faults in sensor networks. They presented an algorithm for generating *irreducible* Identifying Codes in polynomial time. It may be noted that irreducible Identifying Code is only a *minimal* Identifying Code and may not be the *minimum* (or optimal) Identifying Code. In contrast to the algorithm presented in [10], we present an algorithm for construction of optimal Identifying Code for the problem scenario under study. Sen in [11] introduced the novel graph coloring with seepage problem a novel variation of the computation of minimum Identifying Code. Sen in [12] and Basu in [3] analyzed the unique identification of terrorists in terrorist networks. Finally, Basu in [2] studied the monitoring of the health of critical power system equipments by utilizing Identifying Codes.

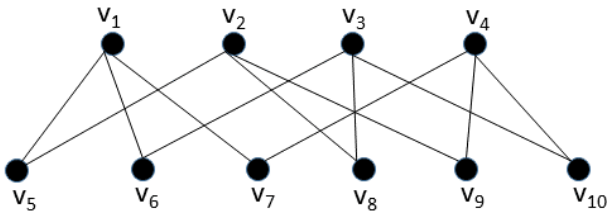


Figure 3: Graph with Identifying Code Set $\{v_1, v_2, v_3, v_4\}$

3 PROPOSED APPROACH

In this report, I use Identifying Code of the *simplest form* and define it as follows. *A vertex set V' of a graph $G = (V, E)$ is defined as the Identifying Code Set (ICS) for the vertex set V, if for all $v \in V$, $N^+(v) \cap V'$ is unique where, $N^+(v) = v \cup N(v)$ and $N(v)$ represents the set of nodes adjacent to v in $G = (V, E)$.* The *Minimum Identifying Code Set* (MICS) problem is to find the Identifying Code Set of *smallest cardinality*. The vertices of the set V' may be viewed as *alphabets* of the code, and the *string* made up with the alphabets of $N^+(v)$ may be viewed as the unique "code" for the node v. For instance, consider the graph $G = (V, E)$ shown in Fig. 3. In this graph $V' = \{v_1, v_2, v_3, v_4\}$ is an ICS as it can be seen from Table 4 that $N^+(v) \cap V'$ is *unique* for all $v_i \in V$. From the table, it can be seen that the code for node v_1 is v_1, the code for v_5 is v_1, v_2, the code for v_{10} is v_3, v_4, etc.

A necessary and sufficient condition for Identifying Codes to exist is that the underlying graph be "twin-free". In other words, this approach fails if there are two nodes $u, v \in V$ such that $N^+(u) = N^+(v)$. In this case, nodes u, v cannot be uniquely identified. In such scenarios, one workaround is to combine nodes u, v to form a super-node.

Table 4: $N^+(v) \cap V'$ results for all $v \in V$ for the graph in Fig. 3

$N^+(v_1) \cap V' = \{v_1\}$	$N^+(v_2) \cap V' = \{v_2\}$
$N^+(v_3) \cap V' = \{v_3\}$	$N^+(v_4) \cap V' = \{v_4\}$
$N^+(v_5) \cap V' = \{v_1, v_2\}$	$N^+(v_6) \cap V' = \{v_1, v_3\}$
$N^+(v_7) \cap V' = \{v_1, v_4\}$	$N^+(v_8) \cap V' = \{v_2, v_3\}$
$N^+(v_9) \cap V' = \{v_2, v_4\}$	$N^+(v_{10}) \cap V' = \{v_3, v_4\}$

`Graph Coloring with Seepage (GCS) Problem`:
The MICS computation problem can be viewed as a novel variation of the classical Graph Coloring problem. I will refer to this version as the *Graph Coloring with Seepage (GCS)* problem. In the classical graph coloring problem, when a color is *assigned* (or injected) to a node, only that node is colored. The goal of the classical graph coloring problem is to use as few distinct colors as possible such that (i) every node receives a color, and (ii) no two adjacent nodes of the graph have the same color. In the GCS problem, when a color is assigned (or injected) to a node, not only does that node receive the color, but also the color *seeps* into all the adjoining nodes. For example, if a node v_i is adjacent to two other nodes v_j and v_k in the graph, then if the color red is injected to v_j, not only v_j will become red, but also v_i will become red as it is adjacent to v_j. Now if the color blue is injected to v_k, not only v_k will become blue, but also

the color blue will seep in to v_i as it is adjacent to v_k. Since v_i was already colored red (due to seepage from v_j), after color seepage from v_k, it's color will be a *combination of red and blue (purple)*. At this point, all three nodes v_j, v_k, and v_i will have "distinct" colors red, blue, and purple, respectively. *The color assigned to a node may be due to: (i) only injection at that node, (ii) only seepage from other adjoining nodes and (iii) a combination of injection and seepage.* The colors injected at the nodes are referred to as *atomic* colors. The colors formed by the combination of two or more atomic colors are referred to as *composite* colors. The colors injected at the nodes (atomic colors) are all *unique*. The goal of the GCS problem is to inject colors to as few nodes as possible, such that (i) every node receives a color, and (ii) no two nodes of the graph have the same color.

Suppose that the node set V' is an ICS of a graph $G = (V, E)$ and $|V'| = p$. In this case if p distinct colors are injected to V' (one distinct atomic color to one node of V'), then by the definition of ICS for all $v \in V$, if $N^+(v) \cap V'$ is unique, all nodes of $G = (V, E)$ will have a unique color (either atomic or composite). Thus computation of MICS is equivalent to solving the GCS problem.

4 METHODOLOGY

The objective of this study is to uniquely determine the source of misinformation propagation in a social network. With billions of users on social networking sites and only a handful of entities (monitors) to verify the veracity of online posts (such as Politifact and Media Bias/Fact Check), it is imperative to have an effective strategy in place, in order to immediately detect the source of such propagation and take swift action against them.

It is evident from the previous section that the mathematical concept of Identifying Code relates to an underlying graph. A social network can be easily represented using a graph $G = (V, E)$, where V denotes the set of users and two users are connected by an edge $e \in E$ if they are friends. The monitors (Politifact or MBFC) should be placed (on the nodes) in a manner such that, if a user (or node) becomes active in propagating misinformation, then they will be uniquely identified. This approach is based on the fact that all the immediate friends or followers of the user initiating the misinformation, will get to know about the misinformation (via Twitter or Facebook posts) . In reality, not all of these friends will participate in the propagation of the misinformation in the next time step. However, in our approach, we assume the contrary, i.e., all the friends (of the initiating user) participate in the propagation of misinformation in the next time step. Thus, the computation of minimum nodes (users) to be monitored, is equivalent to solving the MICS problem. Below, an Integer Linear Program is presented, which solves the MICS problem.

Instance: $G = (V, E)$, an undirected graph, where the node set V denotes the set of users in the social network and there is an edge $e_{ij} \in E$ if i and j are friends.

Problem: Find the smallest subset $V' \subseteq V$, such that injection of colors at these nodes, ensures that each node $v_i \in V$, receives a unique color (either atomic or composite) through seepage.

We use the notation $N(v_i)$ to denote the neighborhood of v_i, for any $v_i \in V$. Corresponding to each $v_i \in V$, we use an indicator variable x_i,

$$x_i = \begin{cases} 1, & \text{if a color is injected at node } v_i, \\ 0, & \text{otherwise} \end{cases}$$

Objective Function:	*Minimize* $\sum_{v_i \in V} x_i$
Coloring Constraint:	$\sum_{v_i \in N(v_j)} x_i \geq 1$, $\forall v_j \in V$
Unique Coloring Constraint:	$\sum_{v_i \in \{N(v_j) \bigoplus N(v_k)\}} x_i \geq 1,\ \forall v_j \neq v_k, \in V$

$N(v_j) \bigoplus N(v_k)$ denotes the Exclusive-OR (symmetric set difference) of the node sets $N(v_j)$ and $N(v_k)$. It may be noted that the objective function ensures that the fewest number of nodes in V are assigned a color. The Coloring Constraint ensures that every node in V receives at least one color through seepage from the colors injected at nodes in V. A consequence of the Coloring Constraint is that, a node in V may receive more than one color through seepage from the colors injected at nodes in V. The Unique Coloring Constraint ensures that, for every pair of nodes (v_j, v_k) in V, at least one node in the node set $N(v_j) \bigoplus N(v_k) \subseteq V$ is injected with a color. This guarantees that v_j and v_k will not receive identical colors through the color seepage from the nodes in V.

5 RESULTS

In this section, I present a set of preliminary results obtained from the Facebook dataset available on SNAP [9]. Table 5 contains the results of the ILP presented in the previous section. In the table, the rows denote the various Facebook networks studied in this effort. All of the graphs under consideration had certain number of twins. As mentioned earlier, one necessary and sufficient requirement for the computation of Identifying Codes is that the graph be "twin-free". To get around this constraint, two nodes which are "twins" can be condensed to form a super node. This formation of a super node ensures that the graph becomes "twin-free" and as a result, an MICS of such a modified graph exists. It may be noted that in such a scenario, the super node receives a unique signature (or identification). To distinguish between the two nodes which formed a super node, deeper investigation is required, which can be accomplished by analyzing the attributes (or behaviour) of the two users separately. This analysis can be done using various data mining techniques.

In the table, the first highlights the results for the "0" graph in the Facebook dataset. Originally, the graph had 333 nodes and 2519 edges. Using a simple twin detection algorithm, we condensed the twin nodes to form super nodes. At the end of this process, it was observed that the number of nodes in the graph decreased to 312, and correspondingly, the number of edges in the graph became 2418. After applying our ILP, we noted that *only* 85 seed nodes (or users) are required to be monitored by entities such as Politifact/MBFC in order for them to have unique identification for all the nodes in the graph (or network). In other words, by monitoring the behaviour of these 85 users, the fact-checking company can uniquely identify

Table 5: Identifying Code Results

Input Graph	Number of Nodes	Number of Edges	Number of Nodes after Twin Removal	Number of Edges after Twin Removal	Number of Nodes to Target	Reduction in Resources
0	333	2519	312	2418	85	**72.75%**
107	1034	26749	1026	2418	125	**87.81%**
348	224	3192	220	3173	40	**81.81%**
414	150	1693	144	1587	32	**77.77%**
686	168	1656	166	1651	30	**81.92%**
698	61	270	56	243	23	**62.26%**
3437	534	4813	517	4686	113	**78.14%**
3980	52	146	46	135	18	**60.87%**

a user if they become active, with respect to propagating misinformation. Once a user becomes active in propagating misinformation, the respective authorities can take corrective measures. Thus for the "0" graph, 85 nodes are sufficient for unique monitoring of 312 nodes in the graph. This signifies a **72.75%** (((312 - 85) / 312) * 100) reduction in computational resources. This is as opposed to monitoring the behavior of all the users in the network. The results for the other graphs are tabulated in Table 5. The average reduction in resources for all the graphs under study turns out to be **75.42%**. This is a significant reduction in resources.

6 FUTURE WORK AND ADVICE SOUGHT

Effective utilization of Identifying Codes will enable fact-checking and social networking companies to uniquely identify the source of misinformation propagation. As a part of my research, I am also currently studying several variants of the MICS problem. The Budgeted MICS problem introduces a limitation on the number of monitors (or sensors) that may be placed in the network. The Targeted MICS problem does not require all the nodes (or users) in the network to have unique signatures, except only the targeted nodes (or users). The Augmented MICS problem assumes that there are already some monitors present in the network, but additional monitors must be adequately placed to enable unique monitoring of all the users. Furthermore, over the past couple of decades, various research groups have studied Identifying Codes and presented approximate algorithms for the same. As a novel contribution, I am trying to leverage the advancements in network embeddings to develop an algorithm for the computation of MICS.

One major assumption of the approach discussed thus far is that the retweeting or sharing misinformation (corresponding to color seepage) is deterministic. In contrast, a user may or may not propagate the misinformation it receives, to its friends. This creates a probabilistic scenario of MICS and I intend to seek insights from the peers and experts at the consortium into this challenge.

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. Arun Sen, for his invaluable inputs and continuous support in my research efforts.

REFERENCES

[1] Mohammad-Ali Abbasi and Huan Liu. 2013. Measuring user credibility in social media. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 441–448.

[2] Kaustav Basu, Malhar Padhee, Sohini Roy, Anamitra Pal, Arunabha Sen, Matthew Rhodes, and Brian Keel. 2018. Health Monitoring of Critical Power System Equipments Using Identifying Codes. In *International Conference on Critical Information Infrastructures Security*. Springer, 29–41.

[3] Kaustav Basu, Chenyang Zhou, Arunabha Sen, and Victoria Horan Goliber. 2019. A Novel Graph Analytic Approach to Monitor Terrorist Networks. *arXiv preprint arXiv:1902.02836* (2019).

[4] Irène Charon, Olivier Hudry, and Antoine Lobstein. 2003. Minimizing the size of an identifying or locating-dominating code in a graph is NP-hard. *Theoretical Computer Science* 290, 3 (2003), 2109–2120.

[5] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.

[6] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. 2017. Fake news mitigation via point process based intervention. *arXiv preprint arXiv:1703.07823* (2017).

[7] Mark G Karpovsky, Krishnendu Chakrabarty, and Lev B Levitin. 1998. On a new class of codes for identifying vertices in graphs. *IEEE Transactions on Information Theory* 44, 2 (1998), 599–611.

[8] Moshe Laifenfeld and Ari Trachtenberg. 2008. Identifying codes and covering problems. *IEEE Transactions on Information Theory* 54, 9 (2008), 3929–3950.

[9] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

[10] Saikat Ray, Rachanee Ungrangsi, De Pellegrini, Ari Trachtenberg, and David Starobinski. 2003. Robust location detection in emergency sensor networks. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, Vol. 2. IEEE, 1044–1053.

[11] Arunabha Sen, Victoria H Goliber, Kaustav Basu, Chenyang Zhou, and Sumitava Ghosh. 2019. On upper and lower bounds of identifying code set for soccer ball graph with application to satellite deployment. In *Proceedings of the 20th International Conference on Distributed Computing and Networking*. ACM, 307–316.

[12] Arunabha Sen, Victoria Horan Goliber, Chenyang Zhou, and Kaustav Basu. 2018. Terrorist Network Monitoring with Identifying Code. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 329–339.

[13] Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 101–102.

[14] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[15] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 430–435.

[16] Jukka Suomela. 2007. Approximability of identifying codes and locating-dominating codes. *Inform. Process. Lett.* 103, 1 (2007), 28–33.

[17] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[18] Ying Xiao, Christoforos Hadjicostis, and Krishnaiyan Thulasiraman. 2006. The d-identifying codes problem for vertex identification in graphs: probabilistic analysis and an approximation algorithm. In *International Computing and Combinatorics Conference*. Springer, 284–298.

Representation Learning for Recommender Systems with Application to the Scientific Literature

Robin Brochier
Université de Lyon, Lyon 2, ERIC EA3083
Digital Scientific Research Technology
Lyon, France
robin.brochier@univ-lyon2.fr

ABSTRACT

The scientific literature is a large information network linking various actors (laboratories, companies, institutions, etc.). The vast amount of data generated by this network constitutes a dynamic heterogeneous attributed network (HAN), in which new information is constantly produced and from which it is increasingly difficult to extract content of interest. In this article, I present my first thesis works in partnership with an industrial company, *Digital Scientific Research Technology*. This later offers a scientific watch tool, *Peerus*, addressing various issues, such as the real time recommendation of newly published papers or the search for active experts to start new collaborations. To tackle this diversity of applications, a common approach consists in learning representations of the nodes and attributes of this HAN and use them as features for a variety of recommendation tasks. However, most works on attributed network embedding pay too little attention to textual attributes and do not fully take advantage of recent natural language processing techniques. Moreover, proposed methods that jointly learn node and document representations do not provide a way to effectively infer representations for new documents for which network information is missing, which happens to be crucial in real time recommender systems. Finally, the interplay between textual and graph data in text-attributed heterogeneous networks remains an open research direction.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; *Artificial intelligence*; • **Information systems** → *Data mining*.

KEYWORDS

representation learning; recommender systems; network embedding; scientific literature

ACM Reference Format:
Robin Brochier. 2019. Representation Learning for Recommender Systems with Application to the Scientific Literature. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314195

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314195

1 PROBLEM

Many applications, which have become everyday tools, offer to search and filter the vast data sources available on the Web. In particular, there is a multitude of platforms dealing with scientific literature. From the simple search engine for scientific articles to the social network for researchers, all use, as data, the daily publications produced around the world. For the researcher facing this deluge of information, it has become difficult, if not impossible, to conduct a regular and exhaustive monitoring of his areas of expertise. The ongoing research presented in this paper, done in partnership with an industrial player [1], deals with the problem of learning representations in heterogeneous networks of documents applied to the recommendation of scientific literature in real time.

If the scientific information of a publication is mainly contained in the text that composes it, rich supplementary information is nested in its metadata. Thus, networks of co-authors, citations and places of publication of an article contain important information for the realization of a scientific recommender system. As such, the scientific literature constitutes a heterogeneous attributed network (HAN) and since new papers are constantly published, this HAN grows in real time (Figure 1 shows an hypothetical example). However, limited information might be observed from the new aggregated nodes. For example, a newly published paper hasn't incoming citation links and a PhD student has a limited number of past co-authorship links. To face this lack of information, the approach considered in the proposed research is to focus on learning strong representations of the attributes, particularly the textual contents of the articles, that can reflect the partially observable underlying network structure.

2 STATE OF THE ART

The quality and informativeness of data representation greatly influence the performance of machine learning algorithms. For this reason, a lot of efforts are devoted to devise new ways of learning representations [1]. In Section 2.1, I describe how the task of learning representations of nodes, *i.e.* network embedding, is tightly connected to word embedding. Then, in Section 2.2, I focus on the interplay between natural language processing and network embedding. Finally, in Section 2.3, I present recent works that extend network embedding techniques to HAN.

[1] *Digital Scientific Research Technology* (*DSRT*) and its web application *Peerus*: https://peer.us.

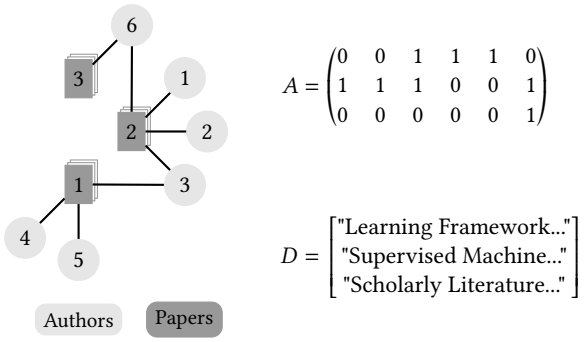


Figure 1: Hypothetical example of scientific literature data, here constituted of 6 authors and 3 papers. The HAN can be noted $G = (V, E, D)$ with V the nodes, E the edges and D the textual content of the papers. We note A the biadjacency matrix of the graph G.

2.1 From Word Embedding to Network Embedding

The distributional hypothesis [20] forms the basis of word embedding algorithms. This assumes that the distributional similarity of words correlates strongly with their semantic similarity. In other words, if we learn a representation of a word that allows us to predict the other words that occurred in its context, we obtain a representation of its meaning.

Skip-Gram [17] is an algorithm that builds representations of words by maximizing the log-likelihood of a multi-set of co-occurring word pairs. Skip-gram with Negative Sampling is a variation proposed in [17] to effectively approach that log-likelihood. This is achieved by reducing the task to a classification which consists in distinguishing pairs of words that co-occur with false pairs that do not co-occur. An alternative approach, GloVe [18], learns representations of words by factoring a matrix of counts of co-occurrences of the words of a corpus. Its objective is to minimize the error of reconstruction of the matrix, considering only the non-zero values of co-occurrence counts.

Even though the distributional hypothesis originated in linguistics and is naturally leveraged for word embedding, Perozzi *et al.* establish the connection with network embedding. To do so, they show that the frequency at which nodes appear in short random walks follows a power-law distribution, like the frequency of words in language [19]. They propose DeepWalk, that consists in applying skip-gram with hierarchical softmax on a corpus of node sequences, deemed equivalent to sentences, generated with truncated random walks [19]. For some specific tasks, the representations learned with DeepWalk offer large performance improvements. Thus, many subsequent works focus on modifying or extending DeepWalk. Node2vec replaces random walks with biased random walks, in order to better balance the exploration-exploitation trade-off, arguing that the added flexibility in exploring neighborhoods helps learning richer representations [7]. VERSE [22] provides a scalable graph embedding algorithm by defining a versatile similarity matrix of the nodes and a learning algorithm using noise-contrastive estimation [8], which provably converges to its objective in contrast to negative sampling.

2.2 Natural Language Processing in Networks of Documents

As a special case of attributed networks, graphs of documents bring together the fields of natural language processing (NLP) and network embedding. A wide variety of unsupervised learning algorithms to represent words and documents have been proposed, from the well-known bag-of-word model [10] to the recently introduced attention-based Transformer [24] adapted for unsupervised pre-training [4]. However, few works have fully explored the interplay between NLP techniques and network embedding.

NetPLSA [16] adapts a topic modelling algorithm by regularizing a statistical topic model with a harmonic regularizer based on a graph structure. It generates topics that reflect the underlying communities of the network, providing cleaner topics than regular statistical models.

In [26], Yang *et al.* prove that skip-gram with hierarchical softmax can be equivalently formulated as a matrix factorization problem. They then propose Text-Associated DeepWalk (TADW), to deal with networks of documents. TADW consists in constraining the factorization problem, with a pre-computed representation of documents via LSA [3]. As such, each node can be represented as the concatenation of a network embedding, and a projected textual embedding. CANE [23] aims to improve node representations in a structured corpus by applying a mutual attention mechanism over the textual contents associated with the vertices of a graph. Given a connected pair of nodes, the model produces textual representations for each node contextually to the other node. In this manner, there are as many representations for a single node as it has neighbors. This model has the advantage to produce interpretable weights for the words of a pair of documents, highlighting those that explain the network structure.

2.3 Heterogeneous Attributed Networks

Real-world networks are often composed of several types of links and nodes which are associated with attributes. For example, the scientific literature is made of articles and authors, with directed citations links between papers, co-authorship links between authors and articles are associated with their textual content and their journal of publication. Many works have extended network embedding to handle heterogeneity and attributes in graphs.

With Metapath2vec [5], the authors propose to operate meta-path based random walks to handle heterogeneous nodes and links. These meta-path are hand-crafted schemes that guide a random walker over the network to generate nodes co-occurrences. Using a similar Skip-Gram based objective as DeepWalk, Metapath2vec achieves significant improvements on multi-class node classification and node clustering over traditional network embedding algorithms. [11] introduces a Label informed Attributed Network Embedding (LANE) framework which jointly projects an attributed network and its labels into a unified embedding space by extracting their correlations. The mapping of the structural proximities in the attributed network and labels into an identical embedding space

via correlation projections produces a significant improvement of the embeddings. Some works go beyond the factorization-based embedding approaches, introducing models that learn a function to generates embeddings by sampling and aggregating features from a node's local neighborhood. GraphSAGE [9] makes use of learnable aggregator functions that allow to infer representations for unseen nodes, given their attributes and links. [25] adapt recent work on attention mechanism to compute a node representation by attending to its neighbors.

3 PROPOSED APPROACH

The proposed approach is intended to produce a novel model for learning representations of nodes and documents in a dynamic heterogeneous network with the goal to compute meaningful recommendations in real-time. The novelty results in the capacity of the model to infer representations of unseen documents for which no network information is available, in the same embedding space as the previously observed nodes. The approach is divided in three steps:

(1) design of a first model to learn node representations that can handle textual attributes. By opposition to TADW, which rely on previously learned LSA representations, this model should learn word and document embeddings from scratch to ease the inference for unseen documents. This step should validate the possibility to learn meaningful text representations from graph information only.
(2) improvement of the model by focusing on natural language processing. The model would be able to predict the similarity of documents in a network, based on their textual content only. It would deal with more advanced NLP techniques and further take advantage of the interplay between word and document representations and the network topology. Compared to CANE, the representations should be produced only from text information (still using the network as training supervision) and a strong emphasis on link prediction for unseen documents should be put.
(3) integration of the heterogeneity to handle different types of nodes and links. This would allow to apply the model to a wider variety of tasks, such as user-item recommendation and expert finding. Handling the diversity of node and link types shouldn't rely on hand-crafted meta-path, as proposed in metapath2vec, but should be learned during the process similarly to GraphSAGE. At this point, the data provided by DSRT from *Peerus* would serve as a strong online evaluation of the proposed model.

Step (1) has been achieved and is detailed in Section 4.2. The results are presented in Section 5.1. More work on its theoretical background will be done in a near future. Step (2) is ongoing research, that I briefly present in Section 4.3 and for which I provide some preliminary results in Section 5.3 motivating the research direction. For all evaluations, I detail the datasets used and the experimental setups in Section 4.1.

4 METHODOLOGY

In this Section, I first provide an overview of the evaluations used for my research, then I detail a contribution corresponding to the first step of my thesis and I finally briefly address the planned methodology for the next steps.

4.1 Evaluation

I first detail some datasets commonly used in the literature. Then I briefly present traditional experiments conducted for evaluating network embedding.

4.1.1 Datasets. I present below two small datasets, Cora and CiteSeer, according to the treatments applied in [21] as well as a larger dataset, DBLP, widely used by the scientific community:

- **Cora** [15] is a network of scientific articles in the field of artificial learning, grouping 7 classes (scientific subdomains) with 2708 documents, 1433 distinct words in the vocabulary and 5429 citations links.
- **CiteSeer** [6] is a network of scientific articles grouping 6 classes on 3312 documents, 3703 distinct words in the vocabulary and 4732 citations links.
- **DBLP** [13] is a database of several million scientific articles in the field of computer science, started in 1993. A powerful disambiguation system makes it possible to identify the authors [14].

Many other non-scientific datasets that present similar data structures exist. Among others, Q&A websites and online encyclopedia provide rich sources of data for which we can tackle similar challenges as with the scientific literature. Moreover, the industrial player supporting these research provides a large dataset of scientific literature with user log activities that allows to apply and evaluate the proposed models to online recommendation tasks.

4.1.2 Experiments. To evaluate network representation learning models, it is common to use the nodes embeddings as input space for a linear algorithm to classify the nodes. For each set of representations produced by a particular algorithm, the proportion of learning representations is varied from 10% to 50% and the average prediction accuracy of the classifier is computed over the rest of the node representations, given a set of ground truth labels. This evaluation was used for the results in Section 5.1.

An extension of this evaluation consists in observing only from 30% to 70% of the nodes when learning the representations. Then, the accuracies of classification are computed on the unobserved nodes, using only their attributes for prediction. As such, we evaluate the algorithm on its capacity to infer representation for new unseen documents. This evaluation was used in Section 5.3. Note that this is different from the inductive prediction performed by GraphSAGE, which also infers representations for unseen nodes, but with the knowledge of the actual new links of these new nodes.

Finally, to evaluate the model of step (2), link prediction constitutes a good evaluation task. Several ways to generate a pair of training/test set exist (random, temporal). The goal is then to distinguish unseen links from non-existing ones. The most suited metric for this is the ROC AUC. The same way as with the previous classification task, it is possible to extend this evaluation for unobserved new documents by hiding a proportion of the nodes (and not of the links) during learning.

4.2 Document Network Embedding

In this section, I present the first contribution of my thesis [2], *GVNR* (Global Vectors for Node Representation), a model to learn node representations with an extension, *GVNR-t*, to handle text-associated nodes. We seek to learn two sets of representations of the nodes $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{n\times d}$, n being the number of nodes in the network and d the dimension of the learned embeddings.

4.2.1 Factorization Problem. We formulate a factorization problem on a random-walk based co-occurrence counts matrix X generated from an input network, measuring the error of reconstruction only for positive coefficients and a fraction of randomly sampled zero coefficients:

$$\underset{U,V,b^U,b^V}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n} s(x_{ij})\Big(u_i \cdot v_j + b_i^U + b_j^V - \log(x_{ij})\Big)^2 \quad (1)$$

b_i^U and b_j^V are two learned biases for the pair of node embeddings. The function s effectively selects the coefficients considered for measuring the reconstruction error:

$$s(x_{ij}) = \begin{cases} 1 & \text{if } x_{ij} > 0, \\ m_i & \text{else, with } m_i \sim \text{Bernoulli}\left(\frac{k\times n_i}{n-n_i}\right). \end{cases} \quad (2)$$

It takes the value 1 for all positive coefficients of X, while for zero coefficients, its value is given by a Bernoulli random variable, m_i. It depends on the number of distinct nodes with which node i co-occurs, n_i. We introduce a global hyper-parameter $k \in \mathbb{N}_*^+$, to control the proportion of zero coefficients incorporated into the reconstruction error.

4.2.2 Extended Model for Networks of Documents. In this brief section, we show how to extend *GVNR* to deal with networks where nodes are short text documents.

Assuming word order is negligible for short documents (such as a scientific abstract), we can model them as bag-of-words and thus represent a document j with a vector $\text{doc}_j \in \mathbb{N}^{+m}$, m being the size of the vocabulary. We can further assume that the meaning of a short text can be captured by averaging the representations of its words [12].

Therefore, with $W \in \mathbb{R}^{m\times d}$ a word embedding matrix, we define the context-vector representation of a node in the following way: $v_j = \frac{\text{doc}_j\, W}{|\text{doc}_j|_1}$

4.3 Improving Document Network Embedding

GVNR-t is able to jointly learn word, document and node embeddings in a network of documents. However, the textual information could highly benefit from more recent works in the field of NLP. In this direction, the recently introduced Transformer as shown great promise in learning dependencies between words for text representation. Besides its low computational complexity and its strong results achieved on neural machine translation and unsupervised pre-training for language understanding, its core unit, the Scaled Dot-Product Attention, provides a good basis for extending *GVNR-t*.

The Scaled Dot-Product Attention takes as input, a set of keys K and values V, corresponding to projected representations of words in a documents, and a query q, possibly any kind of vector lying in the same space as the keys. As output, it generates a weighted sum of the values, whose weights are produced by confronting the keys with the query, following the formula: $\text{Attention}(q, K, V) = \text{softmax}(\frac{qK^T}{\sqrt{d_k}})V$, d_k being the dimension of the query and the keys.

My current research focus on exploring the use of this attention mechanism for mutual attention between pairs of documents in a network. Using pre-trained word embeddings, I try to find a suitable variation of this unit for generating sparse weights (hence using other functions than the softmax) and I explore several ways to build an efficient query q for mutual attention.

5 RESULTS

I first present the results obtained on multi-class classification by *GVNR* and its extension with text and then I show some preliminary results indicating that more emphasis should be set on the representations of the textual content of nodes in a network. Finally, I show an example of a visualization of the weights learned by a preliminary model adapting the Scaled Dot-Product Attention for networks of documents.

5.1 Results for *GVNR* and it Extension With Text

The results presented in Table 1 show the average accuracies for multi-class classification obtained on Cora. First, we observe that *GVNR* produces competitive representations with DeepWalk. Its extension *GVNR-t*, with the integration of the textual content of the documents, significantly improves the quality of the embeddings, achieving even better performances than TADW which relies however on textual representations more time-consuming, obtained with latent semantic analysis.

Table 1: Results of multi-class classification on Cora.

% of training data	10%	20%	30%	40%	50%
DeepWalk	67.8	71.6	74.5	75.8	79.2
GVNR ($x_{\min} = 1$)	**69.5**	**72.6**	**75.9**	**78.1**	**80.2**
DeepWalk+LSA	73.8	77.9	78.4	78.1	78.1
TADW	77.1	78.8	78.2	78.8	78.6
GVNR-t	**79.3**	**80.7**	**80.8**	**81.4**	**81.1**
TADW (text only)	60.5	69.3	72.7	73.6	74.5
GVNR-t (text only)	**74.5**	**76.5**	**78.5**	**78.6**	**79.8**

5.2 Motivation for a Stronger NLP Component

To gain more insight into the textual representations that are learned, Table 1 shows the accuracies obtained by TADW and *GVNR-t* with their respective text components only. We see that *GVNR-t* has significantly higher accuracies than TADW, but it is unclear if this is due to an underlying better natural language understanding. Table 2 shows the results of classification when predicting on unseen documents. We observe that if *GVNR-t* is capable of generalizing on the text attributes of the nodes, TADW relatively fails. However, the results achieved by *GVNR-t* are still lower than expected and motivates the use of more advanced NLP techniques to achieve better generalization.

Table 2: Unseen documents classification accuracies on Cora.

% of training data	30%	40%	50%	60%	70%
TADW	39.7	48.9	50.2	51.5	52.2
GVNR-t	**64.3**	**67.7**	**71.2**	**73.6**	**73.8**

5.3 Mutual Attention for Networks of Documents

My ongoing research is meant to adapt the Scaled Dot-product attention mechanism for network of documents. The hope is to find a way to effectively infer weights for the words of the documents that strongly support (i.e highlight evidence for) the links in the network. Figure 2 shows an example of generated weights with a first draft of such a model. Interestingly, the model highlights words related to the field of reinforcement learning in both texts.

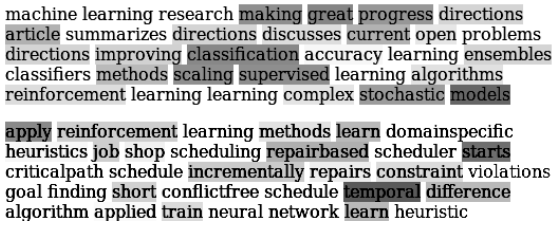


Figure 2: Mutual attention weights for a pair of documents extracted from Cora.

6 CONCLUSION AND FUTURE WORK

Text data and network data are the two most represented information types on the World Wide Web. Building meaningful representations for both is a crucial step for the design of efficient recommender systems. Particularly, the ever growing scientific literature constitute a dynamic heterogeneous text-attributed network. The interplay between the textual content of scientific publications and the networks dynamics of the actors of the research brings strong challenges.

The proposed research aims at discovering an efficient model to represent the variety of nodes and links in the scientific HAN in a unique representation space in order to tackle a wide variety of recommendation tasks. The first works achieved during that research allowed to validate the complementarity of the two sources of information, text and graph, to learn meaningful representations. Ongoing research now aims at improving the natural language understanding component of the model, to truly be able to generate representations for streams of new documents. A last step will focus on extending the coverage of the types of nodes and links that the model can handle and intensively evaluating it on a wide variety of real-world recommendation tasks.

REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[2] Robin Brochier, Adrien Guille, and Julien Velcin. 2019. Global Vectors for Node Representations. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. International World Wide Web Conferences Steering Committee.

[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41, 6 (1990), 391–407.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.

[6] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*. ACM, 89–98.

[7] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864.

[8] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 297–304.

[9] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.

[10] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[11] Xiao Huang, Jundong Li, and Xia Hu. 2017. Label informed attributed network embedding. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 731–739.

[12] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*. 1188–1196.

[13] Michael Ley. 2002. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval*. Springer, 1–10.

[14] Michael Ley. 2009. DBLP: some lessons learned. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1493–1500.

[15] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval* 3, 2 (2000), 127–163.

[16] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 101–110.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[19] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[20] Magnus Sahlgren. 2008. The distributional hypothesis. *Italian journal of linguistics* (2008), 23–53.

[21] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine* 29, 3 (2008), 93.

[22] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. 2018. VERSE: Versatile Graph Embeddings from Similarity Measures. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 539–548.

[23] Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. 2017. Cane: Context-aware network embedding for relation modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1722–1731.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations* (2018). https://openreview.net/forum?id=rJXMpikCZ accepted as poster.

[26] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network representation learning with rich text information.. In *IJCAI*. 2111–2117.

Interactive Visualisation Techniques for the Web of Data

Marie Destandau
Supervised by Emmanuel Pietriga and Caroline Appert
Univ. Paris-Sud, CNRS, Inria
Orsay, France
marie.destandau@inria.fr

ABSTRACT

The RDF format offers powerful possibilities for machines, such as reasoning or federated queries over interlinked datasets. However, presenting RDF data to humans is very challenging: its very structure defeats traditionnal approaches, as it separates information into small pieces, making it difficult for users to make sense of it. My PhD work proposes an approach that presents RDF data in a context, to make them understandable by humans. We first describe S-Paths, a system to support set-based exploration of a dataset's content. We show that it works well on simple models, but that its efficiency is limited by performance issues on very abstract models. Then we lay the basis for a second project, whose aim is to take one more step back and put these sets of entities in a broader context, to give a structural overview of Linked Datasets.

CCS CONCEPTS

• **Information systems** → Presentation of retrieval results; **Browsers**.

KEYWORDS

Visualization, Linked Data, Semantic Web

ACM Reference Format:
Marie Destandau. 2019. Interactive Visualisation Techniques for the Web of Data. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314189

1 PROBLEM

Though RDF format has been designed to be processed by machines, there is a strong need for visualisation and exploration tools: experts involved in publication and reuse of RDF data need to gain a better understanding of their data, and many of the datasets published as RDF would also be of interest to lay users. However, the very structure of RDF makes it difficult to produce efficient visualisations.

The first obvious representation, which accurately reflects the underlying directed-graph structure, is a node-link diagram. Such diagrams are very efficient to describe the data model or a limited number of entities, but they become unreadable over a few dozens of triple statements, while even very small datasets contain several thousands of them.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
TheWebConf, 2019, San Francisco
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314189

Another approach, used by most Linked Data browsers, consists in displaying one page per resource, with all statements directly related to it, as clickable links when they consist of URIs. While this allows users to hop from one resource to another across graphs and datasets, whatever their size, the first drawback is that relevant properties to describe a resource might be several hops away: when the user reaches a piece of information after several clicks, this information is displayed separately from the resource that was originally of interest. Keeping the chain of previous pages in mind requires a significant memorization effort. It is similar to following series of sentences *Peter has a son. The latter has a nephew. The latter has a wife. The latter is a pharmacist.* instead of the single sentence *The nephew of Peter's son is married to a pharmacist.* The mental effort increases as the user browses back and forth to explore other paths. He gathers crumbs of information that he must keep in memory to later make sense of them. However, to display longer chains of information, a browser would need to know where to start and stop. Only browsers with an *a priori* knowledge of the data model are able to do it. A second limitation of this navigation paradigm is that it does not allow to display a resource in the context of other similar resources, whereas this is often an essential feature for sense-making activities.

Tools dedicated to display sets of resources exist. They face the same problems, amplified by the sometimes very large number of entities in a set. In addition, they encounter a problem common to all set-based representations, especially when the range of items is very large: for readability reasons, the relevant information might depend on the number of items in the set (or in the subset in the case of an interactive tool enabling subselections). Thus, considering a collection of 50,000 books, one might want to display the century when they were published, whereas one could list the topics of 500 books, the titles and authors of 50 books, the abstracts and pictures of 10 books, or the full information of a single book. Displaying titles for 50,000 books would simply not make sense because it is not readable.

In this context the first research question for my PhD was: can we use the specificities of Linked Data to automatically identify and display sense-making visualizations at any scale?

2 RELATED WORK

Early visualization tools exposed the raw RDF graph represented as a node-link diagram [9], sometimes using stylesheets to customize the appearance of nodes and links [10]. The approach is generic and accurate, and the resulting representations can be useful to illustrate very small graphs and communicate to a relatively expert audience, but it quickly yields illegible *"big fat graphs"* [12]. In

most cases, the RDF graph is not the right level of abstraction for presentation purposes [5]. It is verbose and the details of how the model structures the data are of low interest to users.

Early linked data browsers approached the problem by producing generic (almost raw) representations of the data in HTML pages, exposing property-value pairs of the current resource of interest as text and clickable URIs. Clicking a URI would attempt to dereference it, and a new HTML page with the properties of the corresponding resource would be displayed. Some browsers featured support for Fresnel [11], a language giving high-level indications about how to display RDF data, which can be used to generate more human-friendly representations of the data. Such languages, however, require that declarative presentation rules be available to the browsers for the different RDF vocabularies involved. Tabulator [1] was one of the first linked data browsers to provide users with more relevant visual representations of RDF data without resorting to such vocabulary-specific presentation languages. However, since Tabulator only features a limited set of views (map, timeline, calendar), it frequently displays data using a generic, more triple-oriented, tree or tabular view on data which are neither spatial nor temporal.

Another significant limitation of many linked data browsers is their strong focus on the *follow-your-nose* browsing strategy. Users explore individual paths from resource to resource over the Web of Data, incrementally fetching the properties of the new resource of interest, without much context (if any) about similar resources. Context information, however, is essential to answering questions related to the distribution of values, to the identification of potential correlations, and to the direct comparison of resources forming a coherent set such as, *e.g.*, all resources of a given type in a particular dataset.

Answering questions such as these, and more generally gaining insights about collections of resources, requires generating multivariate data visualizations (*e.g.*, scatterplots [8]) of a set of properties associated with resources in the set. The Linked Data Visualization Model [4] defines a transformation workflow to dynamically associate one or more datasets with multiple visualizations. Implementations of this workflow typically support a range of data visualization techniques. The methodology defined in [3] focuses on statistical linked data as found in, *e.g.*, RDF data cubes, making it possible to combine data from different linked sources and display them in visualization dashboards. Multiple other projects have been attempting at making it easier for users to generate data visualizations from linked data. Of particular interest here are Visualbox [6] and LinkDaVis [14]. Visualbox enables its users to create different visualizations by writing SPARQL queries and populating predefined visualization templates with the queries' results. The produced visualizations can then be embedded in classic Web pages. LinkDaVis provides users with a hierarchical representation of properties, from which users can choose the ones to visualize. It then performs a heuristic analysis of the data to suggest a ranked list of visualization configurations for these properties, based on different bindings to visual encoding channels. Once a particular visualization is selected, it can further be customized, as in other mixed-initiative approaches such as that adopted in Voyager [15], a tool for the creation of multi-variate data visualizations.

Such tools are very powerful, and enable the creation of a wide variety of visualizations from linked data: maps, timelines, and a range of statistical charts (scatterplot, line plot, density plot, *etc.*). However, they produce standalone visualizations, that can be used on the spot or exported and integrated elsewhere. They do not support browsing over the Web of Data. Indeed, once created, these visualizations feature a very limited level of interactivity. They cannot be used directly to make sub-selections or continue navigation, which requires following links, fetching additional data, and displaying it.

3 PROPOSED APPROACH

Based on the earlier mentioned considerations that the very first condition to enable sense-making activities is readability, and that we would rather follow longer chains of properties than give crumbs of information, our hypothesis is that chains of triple statements in the graph can be used as aggregate steps to reach readable ranges of values. Whatever the number of entities in the set we consider, following chains of triple statements should always lead to a set of values that is either in a readable dimension or can be aggregated.

3.1 S-Paths: Browsing the Content of a Dataset, from Overview to Detail

The first project of this PhD is S-Paths. Its model is based on a *mixed-initiative* approach [14, 15] where the system automatically suggests interesting views for a set of resources, but lets users reconfigure them at will. Exploring a set of resources is an iterative process: whenever users make a selection, S-Paths defaults to what it considers to be the most relevant view on that selection, which users can, again, reconfigure at will. A demo instance is available[1].

3.1.1 Sets of Entities. To identify the initial sets of resources, we relied on the `rdf:type` property. We listed chains of property patterns leading from this set of entities to a set of values. We name these patterns *semantic paths*.

We then defined 6 categories for the paths, corresponding to different possible behaviors in terms of aggregation and display: *datetimes* can be aggregated by years, decades, centuries...; *geographical coordinates* can be aggregated at different scales, corresponding to zooming on a map; *images* can be resized, displayed at different scales, and juxtaposed in grids or mosaic; *numbers* can be aggregated by numerical groups; *text strings* can only be aggregated by similar values, and can be displayed in different layouts depending on their charlength; *URIs* can only be aggregated by similar values, share a common pattern and an approximate similar charlength, and can be used as links. We used these categories, as well as other dimensions listed in Table 1, to characterize the *semantic paths*.

3.1.2 Available Visualizations. Since the conditions of readability are not absolute, but depend on the type of visualization, we relied on a set of views, each one declaring how many dimensions it can display, and for each dimension which categories it can display, and under which conditions (minimum, maximum and optimal). The views are not exclusive, several can apply to a specific case, in this case the optimal conditions will be used to propose the most efficient one.

[1] http://s-paths.net

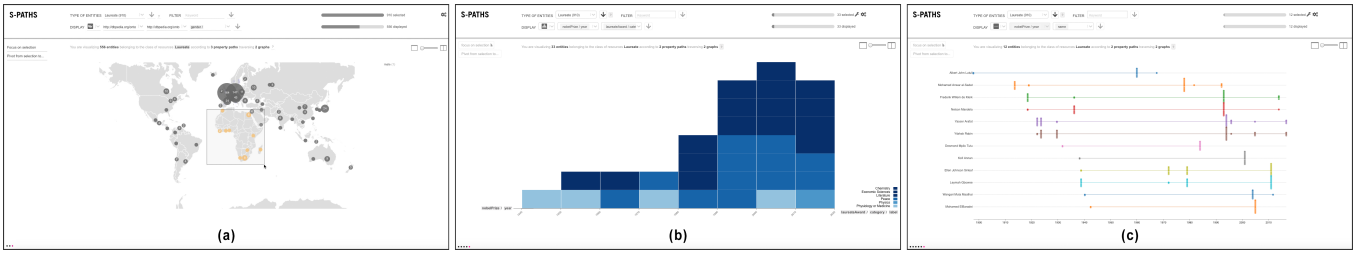


Figure 1: S-Paths can display multiple views on RDF resource sets, showing different properties along paths in the graph. Users can navigate between different resource sets by selecting a subset or pivoting. From (a) a map view of Nobel laureates, selecting those born in Africa and focusing on them yields (b) a histogram showing award year and category (Physics, Literature, *etc.*) for the corresponding laureates. (c) A timeline of events for laureates in the Peace category.

Table 1: Characteristics of semantic paths.

Characteristic	Description
category	one of: `datetime`, `geographical coordinate`, `image`, `number`, `text` or `URI`
depth	number of hops (statements) from the subject to the final property
coverage	percentage of resources in the set for which this path actually exists
count	total number of values for the property at the end of the path, over all resources
unique count	number of unique values for the property at the end of the path, over all resources

S-Paths smoothly animates transitions between views [7] when the two views have entities in common. This provides some basic level of perceptual continuity that contributes to minimizing the cognitive cost of relating the current view to the previous one.

The system also supports the juxtaposition of two consecutive views, as well as brushing and linking between those views: selecting elements in one view immediately highlights the corresponding elements in the other view, further helping users relate views. The same space-filling strategy as above is used to handle brushing & linking between aggregates.

S-Paths also keeps track of all past views and represents them as dots forming a basic timeline displayed in the bottom left corner of the interface. Clicking on one of these dots reverts to the corresponding view, enabling users to easily backtrack.

3.1.3 Matching Algorithm. After having iterated over the collection of views, S-Paths builds a list of optimally-configured views, retaining only the best-scoring paths for each view (Figure 2, 3rd column). It then assigns a score to each view using another normalized weighted average, of:

- *configuration quality*: average of associated path scores;
- *preference*: each type of view has a score which indicates preferences for some types over others based on, *e.g.*, their familiarity or concreteness. As this is subjective and application domain-dependant, these scores can be edited in a configuration file;
- *number of dimensions*: support for more dimensions implies more opportunities for the simultaneous visual mapping of properties.

S-Paths then selects the top-scoring view according to this weighted average, and configures it with the top-ranked semantic paths (Figure 2, 4th column). Lower-ranked paths matching this view can be selected using the *dimensions* menus. The *view* menu lets users switch to any other view compatible with the resource set.

The view reconfiguration capabilities let users change *what dimensions* of resources in the current set are visualized, and *how* they are visualized. Users can also restrict *what resources* to visualize by making direct selections in the currently displayed view: clicking on individual items and aggregates, performing rubber-band selections of contiguous elements, selecting ranges by, *e.g.*, clicking a particular bin on the x-axis of a histogram to select all items in that bar. They can also combine multiple, non-contiguous selections by holding a modifier key (Shift), as in popular graphics-oriented applications such as presentation programs and graphics editors. Once such a sub-selection has been made, users can turn it into the new resource set to explore. The process can be repeated iteratively. Combined with the automatic aggregation of resources along the chosen dimensions, which only occurs when the resource set is too large, this selection mechanism provides users with means to effectively zoom-in on part of the data and get details on demand [13].

3.2 Structural overview of a Linked Dataset

While S-Paths was focused on sets of resources sharing the same type, the second research project of this PhD, explores a broader context, aiming to combine the concept of semantic paths with other pseudo-structural elements, to give an overview of a dataset, and of its links to other datasets.

4 METHODOLOGY

Persona. We started by organizing two workshops, one with 9 Linked Data experts, and another with 7 lay users interested in Linked Data. From these workshop, combined with informal observations during French National Library's hackathon and other Linked Data community events in Paris, we derived 3 persona characters: a Linked Data publisher, a Linked Data reuser, and a lay user interested in the content of Linked Datasets. We used these persona as a leading thread through the development of proofs of concept, to keep in mind relevant user tasks and concerns. We also rely on them to define use cases, and identify relevant users for surveys and studies.

Interviews and Participatory Design. For our second project, we are conducting a series of interviews with experts to better characterize what they need to know about their datasets. We will follow

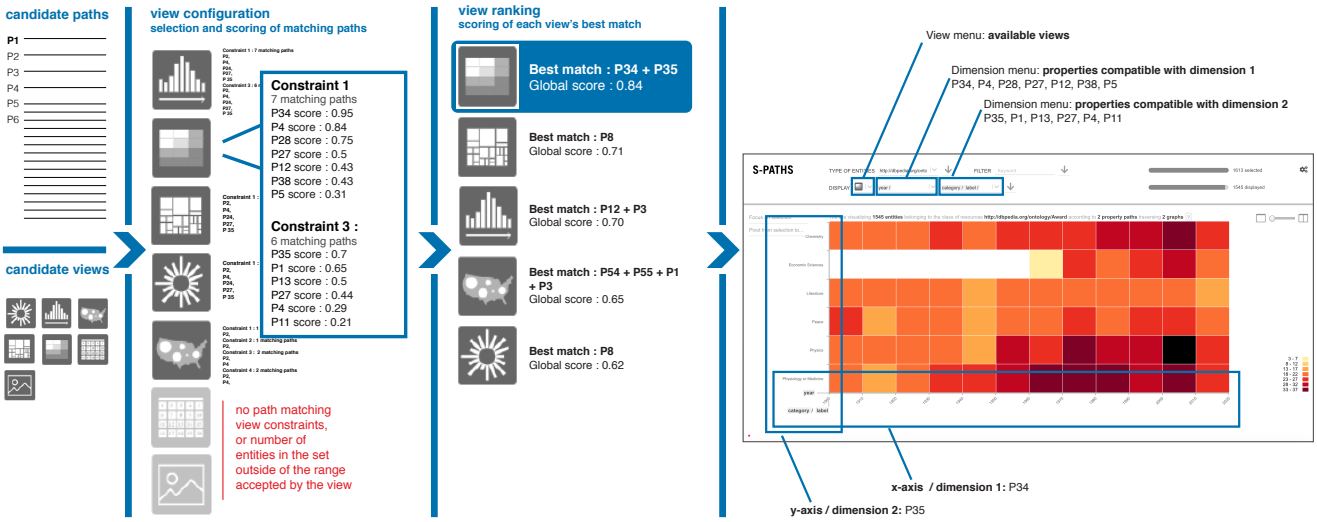


Figure 2: S-Paths' process for generating a default view.

a participatory design approach together with a few of them, who agreed to test our tools and give feedback all along the design and development process. The code will be partly distributed, and run on the endpoints of participants.

Proofs of Concept. We develop high-fidelity prototypes to give proofs of concept of our approach, and evaluate them. Their code is open source. For S-Paths we first made an API to identify and characterize sets of entities and the corresponding semantic paths. Then we started a light version of the visualisation system with two view components, to set up the generic query and transition mechanisms, and then we progressively added other views. The dataset used for development was the Nobel Prize Linked Dataset, which contain 85 797 triple statements, representing 15 classes of ressources. To demonstrate that our tool can work across several graphs, we created a small graph extracted from DBpedia, containing images and geolocations for entities mentioned in the Nobel Dataset, which represents 2 234 triple statements. In a second step we adapted the system to make it work on the French National Library's dataset. Experts at the Bnf extracted a sample that they considered representative, containing 32 106 950 statements. Then we tried the tool on several other datasets from the Linked Data cloud.

Qualitative Evaluation. We are currently running a series of qualitative evaluations, with users representative of each of our persona categories. We demonstrate the tool, and ask users to perform discovery tasks, using a think-aloud protocol. We also collect their feedback in a semi-structured interview. We wish to evaluate if the tool enables to discover new information, and to efficiently understand and remember it.

We are also deploying a version of s-paths at the French National Library to run a longitudinal observation.

5 RESULTS

User Perspective. Though the evaluation of S-Paths is still running, we can already share some observations. Lay users were able to use the tool without knowing about graph databases or what a path in a graph is. However, they were bothered in the navigation by the limitations due to Linked Data: they expected the interface to be more responsive, and complained about not getting previews and rollovers for selections. On the contrary, experts easily understood that such limitations are due to the data's structure and distributed nature, and expressed enthusiasm for the possibility to get overviews. All of them were fully focused on the navigation and 20 minutes appeared very short to them. They would have wanded more time to practice with the tool, and to discover the data. We asked data reusers to explore a dataset as they would have at the start of a hackathon. Our first participant told us that the tool could automatically do most of the work he would usually do when discovering a new dataset. *"When you engage in a hackathon, what you are looking for is irregularity in the data, and this tool finds them and points them out"*. This user noticed that the map showing the co-laureates' affiliation showed surprisingly very few entities in the USA, which revealed that most of the laureates affiliated with US institutions win the prize as a single laureate. This example is particularly interesting, since it would otherwise have demanded an important analysis work from an expert to reach this conclusion, and maybe he would never have had the idea to plot a map with co-laureates only. S-Paths, following the paths, and thus being able to relate geolocations from DBpedia with places across the two graphs, made this starightforward. While developing the tool – using it ourselves and informally demonstrating it to colleagues – we had noticed that most of its visualisations provoke high-level thoughts. For example, a user spotted that there was only one laureate born in the 90's and that she was a female. Another one remarked that there were little shared awards in literature. And indeed, the tool uses aggregation to present the most readable

overview at any step, and, as Bertin states it, *"Useful information comes with cluster"* [2]. Wondering if this would have an impact on memorising and learning, we ended our evaluation with a MCQ. All users – except one, who also reported having no interest in the topic – were able to answer general questions about Nobel prizes (namely, the number of categories, the number of laureates, the time range and the percentage of women), that they would not have been able to answer for the most part before the experiment. A side-effect of the systematic use of aggregation is that it gives the impression of presenting the whole set even when the properties displayed only partially cover that set. Although we clearly indicate coverage in the the top-right corner of the interface (see Figure 1), and even though the tool favors well covered properties, a user pointed out that this should maybe be even more visible.

Scalability and adaptability. So far we tested S-Paths with 7 datasets of various sizes and characteristics: Nobel[2], Data BnF[3], ELI[4], RISM[5], John Peel Sessions[6], Amsterdam Museum[7], Linked Movie DB[8]. Two of them, ELI and RISM did not feature any `rdf:type` statement. Since such statements are the starting point for our tool, we had to generate them in a small adjacent graph according to the model. A limitation of the tool is that it needs to find enough properties in a readable dimensions on the whole set of entities to offer views. At the moment, given the views available and their configuration, this means either dates, numbers, geolocations in any dimensions, or text or URIs with a number of unique values at the end of the path lower than 100. Several classes of resources in the John Peel Sessions dataset did not match this condition. This could be overcome by designing views that group text values and URIs by their first characters(s), thus enabling aggregation for any path. The view would be rated low, because this is not the most meaningful aggregation, but it could serve as an entry point when no other is available.

6 FUTURE WORK

S-Paths, by putting a resource in the context of other similar resources, aims at presenting Linked Data in a manner that is understandable by humans. Continuing with this approach, and taking one more step back, my second research question is: could we identify, analyze and display elements constituting a (pseudo) structure to give an overview of a RDF dataset ?

Indeed, talking with experts during workshops and more informal meetings, we realized that they had difficulties gaining a general understanding of their datasets. In particular, data producers themselves do not seem to have a clear idea about the number of entities in their dataset that are linked to another dataset. Detecting precise links between datasets is not a simple task. If the rest of the analysis can be performed on each SPARQL endpoint independently, links require to query at least two SPARQL endpoints at a time. The support for federated queries and the availability of other endpoints being uncertain parameters, link detection might have to be piloted from a central server. We are currently working on these specific questions.

Finding a way to efficiently detect links between datasets would also be of interest for S-Paths, since it would allow to turn it into a full linked data browser. For now, it is limited to the exploration of multiple graphs hosted behind a single endpoint. Paths traversing datasets could be combined with the pivoting functionality, which already exists in the tool. Pivoting means changing the focus from one set of entities to another along the current path: one can choose to keep the constraints of the current subset or not. This would enable hopping from dataset to dataset. This would imply to find a standard way to communicate the analysis performed in the endpoint.

[2]http://www.nobelprize.org/about/linked-data-examples/
[3]http://api.bnf.fr/dumps-de-databnffrwearecurrentlyworkingona10percentsample
[4]http://data.public.lu/fr/datasets/legilux-journal-officiel-du-grand-duche-de-luxembourg/
[5]https://old.datahub.io/dataset/rism
[6]http://raimond.me.uk/resources/peel.tar.gz
[7]https://bitbucket.org/biktorrr/amlod/downloads/
[8]https://old.datahub.io/dataset/linkedmdb

REFERENCES

[1] Tim Berners-Lee, Yuhsin Chen, Lydia B. Chilton, Dan Connolly, Ruth Dhanaraj, James W. Hollenbach, Adam Lerer, and David Sheets. 2006. Tabulator : Exploring and Analyzing linked data on the Semantic Web.

[2] Jacques Bertin. 1977. *La graphique et le traitement graphique de l'information.* Number 91 (084.21) BER.

[3] Adrian Braşoveanu, Martab Sabou, Arnoa Scharl, Alexandera Hubmann-Haidvogel, and Daniela Fischl. 2016. Visualizing statistical linked knowledge for decision support. *Semantic Web Journal* 8, 1 (2016), 113–137. https://doi.org/10.3233/SW-160225

[4] Josep Maria Brunetti, Sören Auer, Roberto García, Jakub Klímek, and Martin Nečaský. 2013. Formal Linked Data Visualization Model. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS '13)*. ACM, Article 309, 10 pages. https://doi.org/10.1145/2539150.2539162

[5] Aba-Sah Dadzie and Emmanuel Pietriga. 2017. Visualisation of Linked Data - Reprise. *Open Journal Of Semantic Web* 8, 1 (2017), 1 – 21. https://doi.org/10.3233/SW-160249

[6] Alvaro Graves. 2013. Creation of Visualizations Based on Linked Data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS '13)*. ACM, Article 41, 12 pages. https://doi.org/10.1145/2479787.2479828

[7] Jeffrey Heer and George Robertson. 2007. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1240–1247. https://doi.org/10.1109/TVCG.2007.70539

[8] Philipp Heim, Steffen Lohmann, Davaadorj Tsendragchaa, and Thomas Ertl. 2011. SemLens: Visual Analysis of Semantic Data with Scatter Plots and Semantic Lenses. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11)*. ACM, 175–178. https://doi.org/10.1145/2063518.2063543

[9] Emmanuel Pietriga. 2002. IsaViz: a Visual Environment for Browsing and Authoring RDF Models. In *WWW '02: the 11th World Wide Web Conference (Developer's day)*. ACM.

[10] Emmanuel Pietriga. 2006. Semantic Web Data Visualization with Graph Style Sheets. In *Proceedings of the 2006 ACM Symposium on Software Visualization (SoftVis '06)*. ACM, 177–178. https://doi.org/10.1145/1148493.1148532

[11] Emmanuel Pietriga, Christian Bizer, David Karger, and Ryan Lee. 2006. Fresnel: A Browser-independent Presentation Vocabulary for RDF. In *Proceedings of the 5th International Conference on The Semantic Web (ISWC'06)*. Springer-Verlag, 158–171. https://doi.org/10.1007/11926078_12

[12] m. c. schraefel and David Karger. 2006. The Pathetic Fallacy of RDF. In *Semantic Web User Interface Workshop ISWC*.

[13] Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the Symposium on Visual Languages (VL '96)*. IEEE, 336–343. https://doi.org/10.1109/VL.1996.545307

[14] Klaudia Thellmann, Michael Galkin, Fabrizio Orlandi, and Sören Auer. 2015. LinkDaViz – Automatic Binding of Linked Data to Visualizations. In *The Semantic Web - ISWC 2015*. Springer International Publishing, Cham, 147–162.

[15] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Trans. on Visualization and Computer Graphics* 22, 1 (2016), 649–658. https://doi.org/10.1109/TVCG.2015.2467191

Cores and Other Dense Structures in Complex Networks

Edoardo Galimberti
University of Turin & ISI Foundation, Italy
edoardo.galimberti@unito.it

ABSTRACT

Complex networks are a powerful paradigm to model complex systems. Specific network models, e.g., *multilayer networks*, *temporal networks*, and *signed networks*, enrich the standard network representation with additional information to better capture real-world phenomena. Despite the keen interest in a variety of problems, algorithms, and analysis methods for these types of network, the problem of extracting cores and dense structures still has unexplored facets.

In this work, we present advancements to the state of the art by the introduction of novel definitions and algorithms for the extraction of dense structures from complex networks, mainly cores. At first, we define *core decomposition in multilayer networks* together with a series of applications built on top of it, i.e., the extraction of *maximal multilayer cores* only, *densest subgraph* in multilayer networks, the speed-up of the extraction of *frequent cross-graph quasi-cliques*, and the generalization of *community search* to the multilayer setting. Then, we introduce the concept of *core decomposition in temporal networks*; also in this case, we are interested in the extraction of *maximal temporal cores* only. Finally, in the context of discovering polarization in large-scale online data, we study the problem of identifying *polarized communities in signed networks*.

The proposed methodologies are evaluated on a large variety of real-world networks against naïve approaches, non-trivial baselines, and competing methods. In all cases, they show effectiveness, efficiency, and scalability. Moreover, we showcase the usefulness of our definitions in concrete applications and case studies, i.e., the temporal analysis of contact networks, and the identification of polarization in debate networks.

CCS CONCEPTS

• **Mathematics of computing** → **Graph algorithms**; *Graph theory*; *Approximation algorithms*; • **Information systems** → *Social networks*; Data mining.

KEYWORDS

Complex networks; Core decomposition; Dense structures

ACM Reference Format:
Edoardo Galimberti. 2019. Cores and Other Dense Structures in Complex Networks. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314190

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314190

1 PROBLEM

Complex networks model a large variety of systems of high technological and intellectual importance, such as the Internet, the World Wide Web, biological and chemical systems, financial, social, transportation, contact, and communication networks. The need to obtain better understanding of such complex systems summed with their inherent complexity are factors that explain the increasing interest in enhancing complex-networks analytics tools and algorithms. Many network models enrich the standard network representation with additional information that can lead to capturing novel and interesting structural properties of real-world phenomena. For example, *multilayer networks* model complex systems where various types of relations might occur among the same set of entities; *temporal networks* are representations of entities, their relations, and how these relations are established/broken along time; and, *signed networks* represent networked data where edge annotations express whether each edge interaction is friendly (positive) or antagonistic (negative).

Extracting dense structures from large graphs has emerged as a key graph-mining primitive in a variety of scenarios [23], ranging from web mining, to biology, and finance. For instance, dense patterns can help in studying contact networks among individuals to quantify the transmission opportunities of respiratory infections. Anomalously dense patterns among entities in a co-occurrence network have also been used to identify, in real-time, events and buzzing stories. Peculiar dense structures are often associated to polarization and unbalance in online debate networks. Although the literature about complex networks has recently grown extremely fast, the problem of identifying cores and other dense structures in specific types of networks has, surprisingly, still unexplored facets.

Among the many definitions of dense structures, *core decomposition* plays a central role. The *k-core* of a graph is defined as a maximal subgraph in which every vertex is connected to at least k other vertices within such subgraph. The set of all k-cores of a graph G forms the *core decomposition* of G [25]. The importance of core decomposition relies in the fact that it can be computed in linear time [6], and can be used to speed-up/approximate dense-subgraph extraction according to various other definitions. In addition, it has been recognized as an important tool to analyze and visualize complex networks in several domains.

Orthogonal approaches focus on the identification of a single dense portion of a complex network. Several notions of *density* exist in the literature, each of which leads to a different version of the problem. While most variants are **NP**-hard, or even inapproximable, extracting dense subgraphs according to the *average-degree density* (i.e., two times the number of edges divided by the number of vertices) is solvable in polynomial time [14]. As a result, such a density has attracted most of the research in the field, so that the subgraph maximizing the average-degree density is commonly referred to as the *densest subgraph*.

A different bulk of literature deals with methods that look for finding many communities, i.e., dense structures, while partitioning the whole complex network. Among the many approaches, of particular interest is the one provided by the *correlation clustering* framework [4] for signed networks, which asks to partition the vertices into communities so as to maximize (minimize) the number of edges that "agree" ("disagree") with the partitioning, i.e., the number of positive (negative) edges within clusters plus the number of negative (positive) edges across clusters.

In this work, we propose a set of novel definitions of dense structures in complex networks with particular attention on the algorithmic aspects. Specifically, we study (*i*) the problem of *core decomposition in multilayer networks* together with a series of applications built on top of it; (*ii*) the notion of *core decomposition in temporal network*; and, (*iii*) the problem of identifying *polarized communities in signed networks*. The original work about multilayer core decomposition is enclosed in [12]; the enriched version of the same work has been submitted to the TKDD Journal [13]. The work about temporal core decomposition is included in [11], while the one regarding polarized communities is currently under review.

2 STATE OF THE ART

Core decomposition. Given a simple graph $G = (V, E)$, let $d(S, u)$ denote the degree of vertex $u \in V$ in the subgraph induced by vertex set $S \subseteq V$, i.e., $d(S, u) = |\{v \in S \mid (u, v) \in E\}|$.

Definition 1 (Core Decomposition). *The* k-core *(or core of order* k*) of* G *is a maximal set of vertices* $C_k \subseteq V$ *such that* $\forall u \in C_k : d(C_k, u) \geq k$*. The set of all* k*-cores* $V = C_0 \supseteq C_1 \supseteq \cdots \supseteq C_{k^*}$ *(*$k^* = \arg\max_k C_k \neq \emptyset$*) is the* core decomposition *of* G*.*

Core decomposition has been recognized as an important tool to analyze and visualize complex networks [1, 5] in several domains, e.g., bioinformatics, software engineering, and social networks. It has been studied under various settings, such as distributed, streaming/maintenance, and disk-based, and for various types of graphs, such as uncertain, directed, and weighted. The maximal core, i.e., the core of highest order, has been shown to have an important role in determining the propagation patterns of epidemic processes on networks [21].

Azimi-Tafreshi *et al.* [3] study the core-percolation problem in multilayer networks from a physics standpoint without providing any algorithm. They characterize cores on 2-layer Erdős-Rényi and 2-layer scale-free networks, then they analyze real-world (2-layer) air-transportation networks. A type of core decomposition for temporal networks has been proposed by Wu *et al.* [30], who define the (k, h)-core as the largest subgraph in which every vertex has at least k neighbors and at least h temporal connections with each of them. Finally, Zhang *et al.* [31] study the problem of enumerating all maximal cores of a (non-temporal) variant of core decomposition, that turns out to be **NP**-hard.

Densest subgraph. Goldberg [14] provides an exact solution to the densest-subgraph problem based on iteratively solving ad-hoc-defined minimum-cut problem instances. Asahiro *et al.* [2] and Charikar [8] introduce a more efficient (linear-time) $\frac{1}{2}$-approximation algorithm that resembles the one for core decomposition. In fact, it can be proved that the maximal core of a graph is itself a $\frac{1}{2}$-approximation of the densest subgraph.

In the classic definition of densest subgraph there is no size restriction of the output. Variants of the problem with size constraints turn out to be **NP**-hard. A number of works depart from the classic average-degree maximization problem and focus on extracting a subgraph maximizing other notions of density, e.g., quasi-clique-based density, or triangle density. The densest-subgraph problem has also been studied in different settings, such as streaming/dynamic context, and top-k fashion.

Dense structures in multilayer networks. Jethava and Beerenwinkel [19] define the densest common subgraph problem, i.e., find a subgraph maximizing the minimum average degree over all layers of the input graph, and devise a linear-programming formulation and a greedy heuristic algorithm for it. A very recent work by Charikar *et al.* [9] further focuses on the minimum-average and average-minimum formulations by providing several theoretical findings, including **NP**-hardness and hardness of the approximation (for both minimum-average and average-minimum). Jiang *et al.* [20] focus on extracting frequent cross-graph quasi-cliques, i.e., subgraphs that are quasi-cliques in at least a fraction of layers equal to a certain minimum support and have size larger than a given threshold. Interdonato *et al.* [18] are the first to study the problem of local community detection in multilayer networks, i.e., when a seed vertex is given and we want to reconstruct its community by having only a limited local view of the network.

Dense structures in temporal networks. Semertzidis *et al.* [26] introduce the problem of identifying a set of vertices that are densely connected in all or at least k timestamps of a temporal network. The notion of Δ-clique has been proposed in [29], as a set of vertices in which each pair is in contact at least every Δ timestamps. Complementary approaches study the problem of discovering dense temporal subgraphs whose edges occur in short time intervals considering the exact timestamp of the occurrences, and the problem of maintaining the densest subgraph in a dynamic setting. A slightly different yet related body of literature focuses on frequent evolution patterns in temporal attributed graphs, link-formation rules in temporal networks, and the discovery of dynamic relationships and events, or of correlated activity patterns. Bogdanov *et al.* [7] study the problem of finding a subgraph that maximizes the sum of edge weights in a network whose topology remains fixed but edge weights evolve over time.

Balance theory. The concept of signed graph in graph theory appeared in a work by Harary, who was particularly interested in the notion of balance in graphs. Cartwright and Harary generalized Heider's psychological theory of balance in triangles of sentiments to a psychological theory of balance in signed graphs. Harary and Kabell develop a simple linear-time algorithm to test whether a given signed graph satisfies the balance property [15], namely all cycles in the graph contain an even number of negative edges. A more recent line of work develops spectral properties of signed graphs, related to balance theory. Hou *et al.* [16] prove that a connected signed graph is balanced if and only if the least Laplacian eigenvalue is 0. Subsequently, Hou [17] also investigates the relationship between the least Laplacian eigenvalue and the unbalancedness of a signed graph.

Correlation clustering and applications of signed networks. In the original correlation-clustering problem formulation [4], the

number of clusters is not given as input, instead it is part of the optimization. More recent works study the correlation-clustering problem with additional constraints, e.g., a fixed number of clusters or bounds on the cluster sizes. Coleman *et al.* [10] employ the correlation-clustering framework to search for exactly two communities partitioning the whole network.

Signed graphs have also been studied, in different contexts, in machine learning and data mining. Leskovec *et al.* [24] study directed signed graphs and develop status theory, which complements balance theory, to reason about the importance of the vertices in such graphs. Other lines of research include edge and vertex classification, link prediction, community detection, recommendation, and more. A detailed survey on the topic is provided by Tang *et al.* [28].

3 PROPOSED APPROACH

Core decomposition in multilayer networks. A first step towards the identification of novel core/dense structures in complex networks is the study the problem of *core decomposition in multilayer networks*. Let $G = (V, E, L)$ be a multilayer graph, where V is a set of vertices, L is a set of layers, and $E \subseteq V \times V \times L$ is a set of edges. Given an $|L|$-dimensional integer vector $\mathbf{k} = [k_\ell]_{\ell \in L}$, the *multilayer* $\mathbf{k}$*-core* of G is a maximal subgraph whose vertices have at least degree k_ℓ in such subgraph, for all layers $\ell \in L$. Vector $\mathbf{k}$ is dubbed *coreness vector* of that core. The set of all non-empty and distinct multilayer cores constitutes the *multilayer core decomposition* of G. This definition is studied in [3] from a physics standpoint without providing any algorithm. To the best of our knowledge, no prior work has studied how to efficiently compute the complete core decomposition of multilayer networks.

As subsequent step in this direction, we present a series of applications built on top of multilayer core decomposition. First we focus on the problem of extracting only the *maximal multilayer cores*, i.e., cores that are not dominated by any other core in terms of coreness-vector indexes on all the layers. Since maximal cores are orders of magnitude less than all the cores, it is interesting to develop algorithms that effectively exploit the maximality property and extract maximal cores directly, without first computing a complete decomposition. Then, we show how multilayer core decomposition finds application to the problem of *densest-subgraph extraction* from multilayer networks [9, 19]. As a further application, we exploit multilayer core decomposition to speed-up the identification of *frequent cross-graph quasi-cliques* [20]. Finally, we show how multilayer core decomposition can be used to generalize the *community-search* problem [27] to the multilayer setting.

Core decomposition in temporal networks. For temporal networks, we adopt as measure of density of a pattern the *minimum degree* holding among the vertices of a subgraph during a temporal span. The problem of extracting all such patterns is tackled by introducing a notion of *temporal core decomposition* in which each core is associated with its *span*, i.e., an interval of contiguous timestamps, for which the coreness property holds. We call such cores *span-cores*. In all prior works [12, 30], the sequentiality of connections is not taken into account since non-contiguous timestamps can support the same core, then cores cannot be assigned with a clear temporal collocation.

As the total number of time intervals is quadratic in the size of the temporal domain T under analysis, also the total number of span-cores is, in the worst case, quadratic in T, which may be too large an output for human inspection. In this regard, we shift our attention to the problem of finding only the *maximal span-cores*, i.e., span-cores that are not dominated by any other span-core by both the coreness property and the span. Zhang *et al.* [31] also introduce a notion of maximal cores, but theirs cannot be identified in polynomial time (their problem formulation is **NP**-hard) and are not associated with a span.

Polarized communities in signed networks. In the context of studying polarization in large-scale online data, one of the fundamental data-analysis tasks is the *detection of polarization*. We then employ singed networks to study a fundamental problem abstraction for this task, in particular, the problem of *discovering polarized communities*.

In this scenario of application, we propose the problem of finding two communities such that (*i*) within communities there are mostly positive edges while across communities there are mostly negative edges, and (*ii*) the communities are embedded within a large body of other network vertices, which are neutral with respect to the polarization structure. Our hypothesis is that such two-community polarized structure captures accurately controversial discussions in real-world social-media environments. To the best of our knowledge, our work is the first aiming at extracting two hidden polarized communities from signed networks. Our problem formulation deviates from the bulk of the literature where methods typically look for finding many communities while partitioning the whole network. The closest proposal to our problem statement is the work by Coleman *et al.* [10], who employ the correlation-clustering framework and search for exactly two communities, but again they aim at partitioning the whole network.

4 METHODOLOGY

Core decomposition in multilayer networks. A major challenge of computing the complete core decomposition of multilayer networks is that *the number of multilayer cores can be exponential in the number of layers*. In fact, unlike the single-layer case where cores are all nested into each other, no total order exists among multilayer cores. Rather, they form a core lattice defining a relation of partial containment. Within this view, our first contribution is to devise three algorithms that exploit effective pruning rules during the visit of the lattice, thus being much more efficient than a naïve counterpart that computes every core starting from the whole input graph.

Shifting the attention to the problem of computing all and only the maximal cores, a straightforward way of approaching this problem would be to first compute the complete core decomposition, and then filter out the non-maximal cores. However, as the maximal cores are usually much less than the overall cores, it would be desirable to have a method that effectively exploits the maximality property and extracts the maximal ones directly, without computing a complete decomposition. In this work we show that, by means of a clever core-lattice visiting strategy, we can prune huge portions of the search space, thus achieving higher efficiency than computing the whole decomposition.

As a major application of multilayer core decomposition, we focus on the problem of extracting the densest subgraph from a multilayer network. We generalize the problem studied in [9, 19]

by introducing a formulation that accounts for a trade-off between high density and number of layers exhibiting such high density. The importance of the two ingredients of the objective is governed by a parameter β. At first, we prove that the problem is **NP**-hard. Then, we show that computing the multilayer core decomposition of the input graph and selecting the core maximizing the proposed multilayer density function achieves a $\frac{1}{2|L|^\beta}$-approximation for the general multilayer-densest-subgraph problem formulation, and a $\frac{1}{2}$-approximation for the all-layer specific variant studied in [9, 19].

As a further application of our multilayer core-decomposition tool, we show how it can be used as a profitable preprocessing step to speed-up the problem of extracting frequent cross-graph quasi-cliques defined in [20]. Specifically, we prove that the search space of the frequent cross-graph quasi-cliques can be circumstantiated to a number of restricted areas of the input multilayer graph, corresponding to multilayer cores that comply with the quasi-clique conditions. This allows for skipping the visit of unnecessary parts of the input graph, and, thus, speeding-up the whole process, no matter which specific algorithm is used.

Finally, we also provide a generalization of the community-search problem [27] to the multilayer setting, and show how to exploit multilayer core decomposition to obtain optimal solutions to this problem.

Core decomposition in temporal networks. Given a temporal domain T, the number of time intervals included in T is quadratic in the duration of T. Thus, since *a span-core can possibly exist in every time interval*, a $|T|^2$ number of standard core decompositions would in principle be required to compute all the span-cores. As a contribution to the problem of computing all span-cores, we derive containment properties among span-cores, and exploit them to devise an algorithm that computes a span-core decomposition much more efficiently than the naïve approach.

Then, we turn our attention to the problem of finding only the maximal span-cores. As in the multilayer case, a straightforward way of approaching the maximal-span-core-mining problem is to filter out non-maximal span-cores during the execution of an algorithm for computing the whole span-core decomposition. Even in the temporal setting, we can derive a number of theoretical properties about the relationship among span-cores of different temporal intervals and, based on these findings, we show a more efficient way of exploiting the maximality property and extracting maximal span-cores directly, without computing a complete decomposition.

Polarized communities in signed networks. From a theoretical standpoint, we formalize the problem of identifying two polarized communities in a signed network as a *discrete eigenvector* problem. The objective chosen in our formalization is penalized with the size of the solution in order to induce solutions of smaller dimension. Therefore, vertices are only added to one of the two communities if they contribute significantly to the objective. Also, our problem does not enforce balance between the communities. This can be beneficial if there exist polarized communities of significantly different size in the input network.

We prove that our problem formulation is **NP**-hard exploiting a reduction to classic correlation clustering, and propose two intuitive approximation algorithm based on spectral theory. One is deterministic with quality guarantee n, where n is the number of vertices in the graph. It works by simply discretizing the entries of the singular vector $\mathbf{v}$ of the adjacency matrix corresponding to the largest singular value. The other one is randomized with quality guarantee $\sqrt{n}$: it randomly includes each vertex to one of the two communities with probabilities determined by the entries of $\mathbf{v}$. Algorithms running time is essentially the time required to compute the first eigenvector of the adjacency matrix of the input graph. Finally, we propose tweaks to enhance the flexibility of our algorithms and produce a wider variety of results.

5 RESULTS

Core decomposition in multilayer networks. We evaluate our algorithms for mining all multilayer cores and maximal multilayer cores only against their naïve counterparts on real-world multilayer networks in terms of runtime, memory consumption, and search-space exploration. In all cases, the proposed methods result to be more efficient than the baselines, up to four orders of magnitude. We also characterize all multilayer cores and maximal multilayer cores only. As expected, the number of maximal cores is extremely lower: it ranges from the 0.3% to the 22% of the number of all cores.

We also verify the goodness of the proposed applications of multilayer core decomposition. Experimental evidences about our definition of multilayer densest subgraph show that the correct tuning of the parameter β is a key factor in obtaining interesting solutions in real-world networks. Our pruning approach to the extraction of frequent cross-graph quasi-cliques results to be extremely effective (it prunes up to the 99% of the input multilayer graph), being orders of magnitude fastest than the original procedure. Finally, we evaluate the performance of our algorithms for multilayer community search varying the number of query vertices: in all cases they are able to return a solution in a reasonable amount of time, considerably lower than computing the whole decomposition.

Core decomposition in temporal networks. We carry out a comprehensive experimentation on several real-world temporal networks, with millions of vertices, tens of millions of edges, and hundreds of timestamps, to attest efficiency and scalability of our methods. Our algorithms are able to reach up to two orders of magnitude of speed-up with respect to the naïve counterparts.

In order to showcase the usefulness of the proposed definitions, we apply (maximal) span-cores in the analysis of face-to-face contact networks gathered in primary and high schools. Span-cores yield a straightforward temporal analysis of social activities of groups of people within a school day: they are able to identify class hours and breaks, and interesting mixing patterns of students with respect to gender and class. We also analyze the duration of interactions of social groups in schools by studying the distribution of the size of the spans of the maximal span-cores. This analysis confirms a robust statistical behavior and has similar results with respect to simpler characteristics of human interactions, such as the statistics of contact durations. On the basis of this outcome, we provide a simple yet effective procedure to identify anomalous activity patterns in contact networks.

Polarized communities in signed networks. At first, we present a characterization of the polarized communities discovered by our spectral methods that proves the importance of the practical tweaks

and the stability of the randomized solutions. Then, to evaluate the proposed algorithms, we select non-trivial baselines inspired by methods proposed in literature for related problems. Our algorithms produce higher quality solutions, are much faster than the baselines, and can scale to much larger networks. In addition, they are able to identify ground-truth planted polarized communities in synthetic datasets.

A case study about Twitter data regarding the Italian Constitutional Referendum held on December 4, 2016 [22] provides tangible evidence for the goodness of our problem formulation and algorithms in identifying two communities that are polarized about a certain topic. The solution returned by our algorithms almost perfectly split in two communities, favorable and against to the Referendum, the most active users included in the dataset, meaning that the solution identifies the "core" of the controversies, i.e., a set of intensely debating users about the Referendum.

6 CONCLUSIONS AND FUTURE WORK

The problem of extracting cores and other dense structures from complex networks has emerged as a key graph-mining primitive in a variety of scenarios. In this work, we propose novel definitions and methodologies for the identification of such structures in multilayer networks, temporal networks, and signed networks. We show advancements in both theoretical aspects and practical applications. In particular, we study (*i*) core decomposition in multilayer networks together with a series of its applications; (*ii*) core decomposition in temporal networks; and, (*iii*) the identification of polarized communities in a signed network.

This work opens several enticing avenues for further practical applications. In our future investigation about multilayer networks, we plan to employ our core-decomposition tool for the analysis of multilayer brain networks to identify common patterns to patients affected by diseases or under the assumption of drugs and, also, to select features in order to discriminate actual patients from healthy individuals. Similarly, we want to exploit span-cores features for network finger-printing and classification, model validation, and for new ways of visualizing large-scale time-varying graphs. We will also study the role of maximal span-cores with large span in spreading processes on temporal (e.g., contact or presence) networks. The application of the proposed definitions of polarized communities to real-world signed networks can have implications in computational social science problems, e.g., understanding opinion shifts in data streaming from social media sources.

Other future work will deal with more theoretical inquiries. Specifically, we will exploit span-cores for the computation of related notions, such as *community search* or *densest subgraph* in temporal networks. We also plan to study the tightness of our analysis of approximation of the two algorithms for the extraction of polarized communities from signed networks, and derive heuristic algorithms that return meaningful solutions in practice; finally, it would be interesting to extend this problem to detect an arbitrary number of communities.

REFERENCES

[1] J Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. 2006. Large scale networks fingerprinting and visualization using the k-core decomposition. In *Advances in neural information processing systems*. 41–50.

[2] Yuichi Asahiro, Kazuo Iwama, Hisao Tamaki, and Takeshi Tokuyama. 2000. Greedily finding a dense subgraph. *Journal of Algorithms* 34, 2 (2000), 203–221.

[3] N Azimi-Tafreshi, J Gómez-Gardenes, and SN Dorogovtsev. 2014. k- core percolation on multiplex networks. *Physical Review E* 90, 3 (2014), 032816.

[4] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine learning* 56, 1-3 (2004), 89–113.

[5] Vladimir Batagelj, Andrej Mrvar, and Matjaz Zaversnik. 1999. Partitioning Approach to Visualization of Large Graphs. In *Int. Symp. on Graph Drawing*. 90–97.

[6] Vladimir Batagelj and Matjaž Zaveršnik. 2011. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification* 5, 2 (2011), 129–145.

[7] Petko Bogdanov, Misael Mongiovì, and Ambuj K Singh. 2011. Mining heavy subgraphs in time-evolving networks. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 81–90.

[8] Moses Charikar. 2000. Greedy approximation algorithms for finding dense components in a graph. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*. Springer, 84–95.

[9] Moses Charikar, Yonatan Naamad, and Jimmy Wu. 2018. On Finding Dense Common Subgraphs. *arXiv preprint arXiv:1802.06361* (2018).

[10] Tom Coleman, James Saunderson, and Anthony Wirth. 2008. A local-search 2-approximation for 2-correlation-clustering. In *European Symposium on Algorithms*. 308–319.

[11] Edoardo Galimberti, Alain Barrat, Francesco Bonchi, Ciro Cattuto, and Francesco Gullo. 2018. Mining (maximal) span-cores from temporal networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 107–116.

[12] Edoardo Galimberti, Francesco Bonchi, and Francesco Gullo. 2017. Core Decomposition and Densest Subgraph in Multilayer Networks. In *CIKM 2017*.

[13] Edoardo Galimberti, Francesco Bonchi, Francesco Gullo, and Tommaso Lanciano. 2018. Core Decomposition in Multilayer Networks: Theory, Algorithms, and Applications. *CoRR* abs/1812.08712 (2018). arXiv:1812.08712 http://arxiv.org/abs/1812.08712

[14] A. V. Goldberg. 1984. *Finding a Maximum Density Subgraph*. Technical Report. University of California at Berkeley.

[15] Frank Harary and Jerald A Kabell. 1980. A simple algorithm to detect balance in signed graphs. *Mathematical Social Sciences* 1, 1 (1980), 131–136.

[16] Yaoping Hou, Jiongsheng Li, and Yongliang Pan. 2003. On the Laplacian eigenvalues of signed graphs. *Linear and Multilinear Algebra* 51, 1 (2003), 21–30.

[17] Yao Ping Hou. 2005. Bounds for the least Laplacian eigenvalue of a signed graph. *Acta Mathematica Sinica* 21, 4 (2005), 955–960.

[18] Roberto Interdonato, Andrea Tagarelli, Dino Ienco, Arnaud Sallaberry, and Pascal Poncelet. 2017. Local community detection in multilayer networks. *Data Min. Knowl. Discov.* 31, 5 (2017), 1444–1479.

[19] Vinay Jethava and Niko Beerenwinkel. 2015. Finding dense subgraphs in relational graphs. In *ECML-PKDD*. 641–654.

[20] Daxin Jiang and Jian Pei. 2009. Mining frequent cross-graph quasi-cliques. *TKDD* 2, 4 (2009), 16.

[21] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. 2010. Identification of influential spreaders in complex networks. *Nature physics* 6, 11 (2010), 888.

[22] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance Evolution and Twitter Interactions in an Italian Political Debate. In *International Conference on Applications of Natural Language to Information Systems*. 15–27.

[23] Victor E. Lee, Ning Ruan, Ruoming Jin, and Charu C. Aggarwal. 2010. A Survey of Algorithms for Dense Subgraph Discovery. In *Managing and Mining Graph Data*.

[24] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1361–1370.

[25] S. B. Seidman. 1983. Network structure and minimum degree. *Social Networks* 5, 3 (1983), 269–287.

[26] Konstantinos Semertzidis, Evaggelia Pitoura, Evimaria Terzi, and Panayiotis Tsaparas. 2016. Best Friends Forever (BFF): Finding Lasting Dense Subgraphs. *arXiv preprint arXiv:1612.05440* (2016).

[27] Mauro Sozio and Aristides Gionis. 2010. The Community-search Problem and How to Plan a Successful Cocktail Party. In *KDD*. 939–948.

[28] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. A survey of signed network mining in social media. *ACM Computing Surveys (CSUR)* 49, 3 (2016), 42.

[29] Tiphaine Viard, Matthieu Latapy, and Clémence Magnien. 2016. Computing maximal cliques in link streams. *Theoretical Computer Science* 609 (2016), 245–252.

[30] Huanhuan Wu, James Cheng, Yi Lu, Yiping Ke, Yuzhen Huang, Da Yan, and Hejun Wu. 2015. Core decomposition in large temporal graphs. In *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 649–658.

[31] Fan Zhang, Ying Zhang, Lu Qin, Wenjie Zhang, and Xuemin Lin. 2017. When engagement meets similarity: efficient (k, r)-core computation on social networks. *Proceedings of the VLDB Endowment* 10, 10 (2017), 998–1009.

Extracting actionable information from Security Forums

Joobin Gharibshah
jghar002@ucr.edu
University of California Riverside
Riverside, California

ABSTRACT

The goal of this work is to systematically extract information from hacker forums, whose information would be in general described as unstructured: the text of a post is not necessarily following any writing rules. By contrast, many security initiatives and commercial entities are harnessing the readily public information, but they seem to focus on structured sources of information. Here, we focus on the problem of analyzing text content in security forums. A key novelty is that we use user profiles and contextual features along with transfer learning approach and also embedding space to help us identify and refine information that we could not get from security forum with trivial analysis. We collect a wealth of data from 5 different security forums. The contribution of our work is twofold; (a) we develop a method to automatically identify through the forums malicious IP addresses (b) we also propose a systematic method to identify and classify user-specified threads of interest into four categories. We further showcase how this information can inform knowledge extraction from the forums. As the cyber-wars are becoming more intense, having early accesses to useful information becomes more imperative to remove the hackers first-move advantage, and our work is a solid step towards this direction.

KEYWORDS

Security forums, Text analysis, Word embedding, Transfer learning

ACM Reference Format:
Joobin Gharibshah. 2019. Extracting actionable information from Security Forums. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3314197

1 INTRODUCTION

Security forums hide a wealth of information, but mining it requires novel methods and tools. The problem is driven by practical forces: there is useful information that could help improve security, but the volume of the data requires an automated method. The challenge is that there is a lot of "noise", there is lack of structure, and an abundance of informal and hastily written text. At the same time, security analysts need receive focused and categorized information, which can help their task of shifting through it further. Here, we focus on a specific question. In particular, we want to extract as much useful information from hacker/security forums as possible in order to perform (possibly early) detection of potential malicious content such as IP addresses and threads of interest for the security analysts. In this research we are looking to analyze the text in order to identify and classify content of interests. Twofold of the interest here are IP addresses and user-specified security discussion threads. For example threads describing hacking tutorials or announcing emerging attacks.

In the first fold we identify and characterize IP addresses mentioned in text of security forums. The problem that we address here is to find all the IP addresses that are being reported as malicious in a forum. In other words, the input is all the posts in a forum and the expected output is a list of malicious IP addresses.

Interestingly as we showed in our first research [4, 5], not all of the reported IP addresses are malicious, which makes the classification necessary. It turns out that this is a two-step problem. First, we need to solve the IP Identification problem: distinguishing IP addresses from other numerical entities, such as a software version. Second, we need to solve the IP Characterization problem: characterizing IP address as malicious or benign. The extent of the Identification problem caught us by surprise: we find 1820 non-address dot-decimals,

As its key novelty, our approach by utilizing a simple transfer learning technique, minimizes the need for human intervention. First, once initialized with a small number of security forums, it does not require additional training data to mine new forums. Second, it addresses both the Identification and Characterization problems [6].

In the second fold, we propose a systematic approach to identify and classify threads of interest based on an embedding approach. We consider two associated problems that together provide a complete solution to this problem.

First, the input is all the data of a forum, and the user specifies its interest by providing one or more bag-of-words of interest. Arguably, providing keywords is a relatively easy task for the user. The goal is to return all the threads that are of interest to the user, and we use the term "relevant" to indicate such threads. We use the term Identification to refer to this problem. A key challenge here is how to create a robust solution that is not overly sensitive to the omission of potentially important keywords.

Second, we add one more layer of complexity to the problem. To further facilitate the user, we want to group the relevant threads into classes. We utilize the embedding domain which captures the similarity and the context of word to represent in multi-dimensional space. We refer to this step as the Characterization problem. Given a security forum, we want to extract threads of interest to a security analyst.

2 RELATED WORK

We summarize related work clustered into areas of relevance.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314197

a. Extracting and classifying entities from security forums. Recently there have been a few efforts focused on extracting entities of interest in security forums. There is a main efforts that study IP addresses and security forums [3]. In which, they focus on the spatiotemporal properties of Canadian IP addresses in forums without employing any identification and classification methods, which are the focus of our work. Various efforts have attempted to extract other types of information from security forums. A few recent studies identify malicious services and products in security forums by focusing on their availability and price [13, 15]. Another interesting work [18] uses a word embedding technique focusing identifying vulnerabilities and exploits. Other efforts study the users of security forums, group them into different classes, and identify their roles and social interactions [7, 16, 21].

b. Transfer learning. There is extensive literature on transfer learning [1, 2] and several good surveys [20], which inspired our approach. However, to the best of our knowledge, we have not found any work that address the same domain-specific challenges or uses all the steps of our approach, which we described in 3

c. NLP, Bag-of-Words, and Word Embedding techniques. Natural Language Processing is a vast field, and even the more recent approaches, such as word embedding have benefited from significant numbers of studies [10, 12, 17, 19].

3 METHODOLOGY

In this section we explain the proposed approaches to address each fold of our research.

3.1 Analyzing IP addresses

Here we present two problems regarding IP addresses in the security forums and explain our solutions to tackle them.

3.1.1 IP addresses Identification Problem. We describe our proposed method to identify IP addresses in the forum.

The IP address format. The vast majority of IP addresses in the forums follow the *IPv4* dot-decimal format, which consists of 4 decimal numbers in the range [0-255] separated by dots. We can formally represent the dot-decimal notation as follows: *IPv4* $[x_1.x_2.x_3.x_4]$ with $x_i \in [0-225]$, for $i = 1, 2, 3, 4$. Note that the newer *IPv6* addresses consists of eight groups of four hexadecimal digits, and our algorithms could easily extend to this format as well. For now, they are out of the scope of this research.

The challenge: the dot-decimal format is not enough. If IP addresses were the only numerical expressions in the forums with this format, the Identification problem could have been easily solved with straightforward text processing and Named-Entity Recognition (NER) tools, such as the Stanford NER models. However, there is a non-trivial number of other numerical expressions, which can be misclassified as addresses such as software version and system logs.

To this end, we propose a method to solve the IP Identification problem, a supervised learning algorithm. We first identify the features of interest as we discuss below. We then train a classifier using the Logistic Regression method gives the best results among the several methods using 10-fold cross validation on our ground-truth as we described in the previous section.

Feature selection.

We use three sets of features in our classification.

a. Contextual information (*TextInfo*): Inspired by how a human would determine the answer, we focus on the words surrounding the dot-decimal structure. For example, the words *"server"* or *"address"* suggests that the dot-decimal is an address, while the words *"version"* or a software name, like *"Firefox"* suggests the opposite. We consider the frequency of the surrounding words, before and after the dot-decimal structure, in our classification.

b. The numerical values of the dot-decimal (*DecimalVal*): We use the numerical value of the four numbers in the the dot-decimal structure as features. The rationale is that non-addresses, such as software versions, tend to have lower numerical values. This insight was based on our close interaction with the data.

c. The combined set (*Mixed*): We combine the two feature sets to create in order to leverage their discriminating power.

We saw that using *Mixed* outperforms two other feature sets.

3.1.2 IP addresses Characterization Problem. We develop a supervised learning algorithm to characterize IP addresses. Here, we assume that we have labeled data, and we discuss how we handle the absence of ground truth in section 3.1.3. We first identify the appropriate set of features which we discuss below. We then train a classifier.

Features sets for the Characterization problem. We consider and evaluate two sets of features in our classification.

a. Text information of the post : *PostText* We use the words and their frequency of appearance in the post. Here, we use the TF-IDF technique again to better estimate the discriminatory value of a word by considering its overall frequency.

b. The Contextual Information set: *ContextInfo* . We consider an extended feature set that includes both the *PostText* features, but also features of the author of the post. These features capture the behaviour of the author, including frequency of posting, average post length etc. These features were introduced by our earlier work [4], with the rationale that profiling the author of a post can help us infer their intention and role and thus, improve the classification.

We saw that, by using *PostText* features on their own, we obtain slightly better results.

3.1.3 Transfer Learning with Cross-Seeding for Identification and Characterization . In both introduced problems, we face the following conundrum:

a. the classification efficiency is better when the classifier is trained with forum-specific ground-truth, but,

b. requiring ground-truth for a new forum will introduce manual intervention, which will limit the practical value of the approach.

We propose to do cross-forum learning by leveraging transfer learning approaches [2, 14]. We use the terms *source* and *target* domain to indicate the two forums with the target forum not having ground-truth available.

We propose an algorithm that will help us develop a new classifier for the target forum by using the old classifier to create training data as we explain below.

Our Cross-Seeding approach. We propose to create training data for the target forum following the four steps below.

a. Domain adaptation. The main role of this step is to ensure that the source classifier can be applied to the target forum. The main issue in our case is that the feature sets can vary among forums.

Recall that, for both classification problems, we use the frequency of words and these words can vary among forums. We adopt an established approach that works well for text classification [2]: we take the union of the feature sets of the source and target forums. The approach seems to work sufficiently well in our case, as we see later.

b. Creating seed information for the target forum. Having resolved any potential feature disparities, we can now apply the classifier from the source forum to the target forum.

We create the seeding data by selecting instances of the target domain, for which the classification confidence is high. Most classification methods provide a measure of confidence for each classified instance and we revisit this issue in section 4.

c. Training a new classifier for the target forum. Having the seed information, we train the classifier directly.

d. Applying the new classifier on the target forum. In this final step, we apply our newly-trained forum-specific classifier on the target forum.

3.2 Analyzing threads of interest

We describe our approach toward analyzing the user-specified thread of interest first by identifying then classifying such thread.

3.2.1 Identifying threads of interest. We present our approach for selecting relevant threads starting from sets of keywords provided by the user. Our approach consists of the following phases: (a) a keyword matching phase, where we use the user-defined keywords to identify relevant threads that contain these keywords, and (b) a similarity-based phase, where we identify threads that are "similar" to the ones identified above. The similarity is established at the word embedding space as we describe later.

Phase 1: Keyword-based selection

Given a set or sets of keywords, we identify the threads where these keywords appear. A simple text matching approach can distinguish all occurrence of such keywords in the forum threads. In more detail, we follow the steps below:

Step 1: The user provide a set or sets of keywords , which capture the user's topics of interest. Having sets of keywords enables the user to specify combinations of concepts. For example, in our case we use, the following sets: (a) hacking related, (b) exhibiting concern and agitation, and (c) searching and questioning.

Step 2: We count the frequency of each keyword in all the threads. This can be done easily with elastic search or any other straightforward implementation.

Step 3: We identify the relevant threads, as the threads that contain a sufficient number of keywords from each set of keywords. This can be defined by a threshold for each set of keywords.

Phase 2: Similarity-based selection

We propose an approach to extract additional relevant threads based on their similarity to existing relevant threads. In following steps, in which input is a forum, a set of keywords, and set of relevant threads, as identified by the keyword-based phase above.

Step 1. Determining the embedding space. We project every word as a point in a m-dimensional space using a word embedding approach[12]. Therefore, every word is represented by a vector of m dimensions.

Step 2. Projecting threads. We project all the threads in an appropriately constructed multi-dimensional space: both the relevant threads selected from the keyword-based selection and the non-selected ones. The thread projection is a function of the vectors of its words and captures both the average and the maximum values of the vectors of its words[17].

Step 3. Identifying relevant threads. We identify more relevant threads among the non-selected threads that are "sufficiently-close" to the relevant threads in the thread embedding space with cosine similarity measure.

The advantage of using similarity at the level of threads is that thread similarity can detect high-order levels of similarity, beyond keyword-matching. Thus, we can identify threads that do not necessary exhibit the keywords, but use other words for the same "concept"

3.2.2 Classifying threads of interest. We present our approach for classifying relevant threads into user-defined classes based on embeddings representation of words and also contextual features.

Given a thread, we calculate its projection in the embedding space based on the method proposed in the previous section. The embedding approaches have been used in the task of the text classification recently.

Using contextual features.

Apart from the words and their embeddings in the forum, we can also consider other types of features, which we refer to as contextual features of the threads. One could think of various such features, but here we list the features that we use in our evaluation: (1) number of newlines, (2) length of the text, (3) number of replies in the thread (following posts after the first post), (4) average number of newlines in replies, (5) average length of replies, and (6) the aggregated frequency of the words of each bag-of-words set provided by the user.

These features capture contextual properties of the posts in the threads, and provide additional information not necessarily captured by the words in the thread.

Empirically, we find that these features improve the classification accuracy significantly. The inspiration to introduce such features came from manually inspection of posts and threads. For example, we observed that Hacks and Experiences usually have longer posts than other. Moreover, Hacks threads contain a larger number of newline characters. An interesting question is to assess the value of such metrics when used in conjunction with word-based features.

4 EXPERIMENTAL RESULTS

4.1 Data Crawling and Labeling

We have collected data from five different forums, which cover a wide spectrum of interests and intended audiences. We present basic statistics of our forums in Table 1. We have developed an efficient and customizable python-based crawler, which can be used to crawl online forums, and it could be of independent interest.

For validating our classification method, we labeled data for each fold of the research as following:

In the IP addresses analysis, For the Identification problem, we could not find any external sources of information and benchmarks. To establish our ground-truth, we selected dot-decimal expressions

	WildersSec.	OffensComm.	HackThisSite	EthicalHackers	Darkode
Posts	302710	25538	84125	54176	75491
Threads	28661	3542	8504	8745	7563
Users	14836	5549	5904	2970	2400
Dot-decimal	4325	7850	1486	1591	1097
IP found	3891	6734	1231	1330	1082

Table 1: The basic statistics of our forums

uniformly randomly, and we used four different individuals for the labelling.

For the Characterization problem, we make use of the VirusTotal site which maintains a database of malicious IP addresses by aggregating information from many other such databases. We also provide a second level of validation via manual inspection.

In the thread of interest analysis, we need groundtruth to do both the training and the validation. We randomly selected 450 among the relevant threads from each forum as selected by the identification part. The labelling involved three manual evaluations based on the definitions and examples of the four classes, which we listed above.

4.2 Results in IP addresses analysis

We evaluate our approach focusing on the performance of Cross-Seeding for both the Identification and the Characterization problems.

We use Logistic Regression as our classification engine, which performed better than several others, including SVM, Bayesian networks, and K-nearest-neighbors. In Cross-Seeding , we use the Logistic Regression's prediction probability with a threshold of 0.85 to strike a balance between sufficient confidence level and adequate number of instances above that threshold.

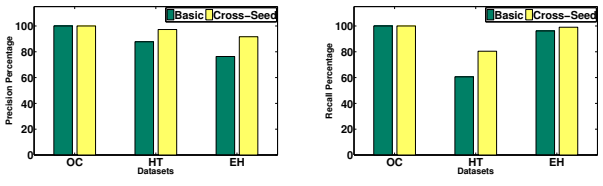
A. The IP Identification problem. In this problem, our classification approach exhibits 98% precision and 96% recall on average across all our sites, when we train with ground-truth for each forum.

Identification : 95% precision with Cross-Seeding and out performs Basic . We show that our cross-training approach is effective in transferring the knowledge between domains. We use the classifier from WildersSecurity and we use it to classify three of the other forums, namely, OffensiveCommunity , EthicalHackers , and HackThisSite .

In figure 1, we show the results for precision and recall of cross-training using Basic (use the classifier from the source forum on the target forum) and Cross-Seeding . We see that Cross-Seeding improves *both* precision and recall significantly. For example, for HackThisSite , Cross-Seeding increases the precision from 57% to 79% and the recall from 60% to 78%. Cross-Seeding improves the precision by 8% and recall by 7% on average for the experiment shown in figure 1. The average precision increased from 88% to 95% and the average recall increased from 85% to 97%.

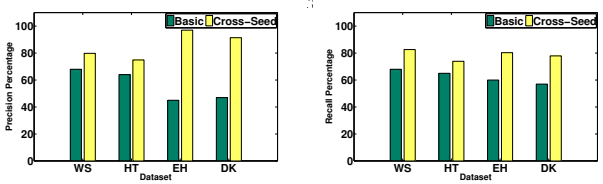
B. The IP Characterization problem. We evaluate our approach for solving the Characterization problem without per-forum training data. We can achieve 93% precision and 92% recall on average across all the forums, when we train with ground-truth for each forum.

a. Characterization : 88% precision on average with Cross-Seeding and outperforms Basic . Using OffensiveCommunity as source, and we classify WS , HackThisSite , and Darkode as shown



(a) Precision (b) Recall

Figure 1: Identification : Cross-Seeding improves both Precision and Recall. Using WildersSecurity to classify



(a) Precision (b) Recall

Figure 2: Characterization : Cross-Seeding improves both Precision and Recall. Using OffensiveCommunity as source.

in figure 2. Our Cross-Seeding approach can provide 88% precision and 82% recall on average.

we show thatby using OffensiveCommunity as our source, we see that Cross-Seeding improves the precision by 28% and recall by 16% on average across the forums compare to the Basic approach. We also observe that the improvement is substantial: Cross-Seeding improves both precision and recall in all cases.

b. Using more source forums improves the Cross-Seeding performance significantly. We quantify the effect of having more than one source forums in the classification accuracy of a new forum. We use and WildersSecurity as our training forums, and we use Cross-Seeding for OffensiveCommunity , HackThisSite , and Darkode . First, we use the source forums one at a time and then both of them together. We evaluated the average improvement of having two source forums over having one for each target website. Using two source forums increases the classification precision by 13% and the recall by 17% on average.

4.3 Results in threads of interest analysis

We present our experimental results and evaluation of our approach in analyzing threads of interest. We use the three forums (OffensiveCommunity , HackThisSite ,) that presented in Table 1 and the groundtruth, which we created as we explained in section 4.1.

Keywords sets: We considered three keyword sets to capture relevant threads. These keywords set are: (a) hacking related, (b) exhibiting concern and agitation, and (c) searching and questioning. We collected a set of 300 keywords in three sets. We started with a small core group of keywords, which we expanded by adding their synonyms using thesaurus.com and Google's dictionary.

Embedding parameters: We set the window size to 10 and we tried several different values as the dimension of the embedding between 50-300, and we found that $m = 100$ with the highest accuracy value.

Similarity threshold: The similarity threshold determines the "selectiveness" in identifying similar threads, as we described in a previous section. We find that a value of 0.96 worked best among all the different values we tried.

Our classifier. We use random forest as our classification engine, which performed better than several others that we examined, including SVM, Neural Networks, and K-nearest-neighbors. Results are not shown due to space limitations.

Baseline methods. We evaluate our approach against three other state of the arts methods, which we briefly describe below. **Bag of Words (BOW)**: This methods users the word frequency (more accurately the TFIDF value) as its main feature [4, 8]. **Non-negative Matrix Factorization (NMF)**: This method users linear-algebra to represent high-dimensional data into low-dimensional space, in an effort to capture latent features of the data [11]. **FastText (FT)**: There is a family of methods that use the word2vec as their basis, and use a recently proposed method [9].

We present the results of our proposed methods.

Our similarity-based method is robust to the number of initial keywords. We evaluate the impact of the number of keywords to the similarity based method. In Figure 3, we show the robustness of each identification methods to the initial set of keywords for OffensiveCommunity . By adding 60 keywords, from 240 to 300, the keyword-based method identifies 25% more threads, while the similarity based method has only 7% increment.

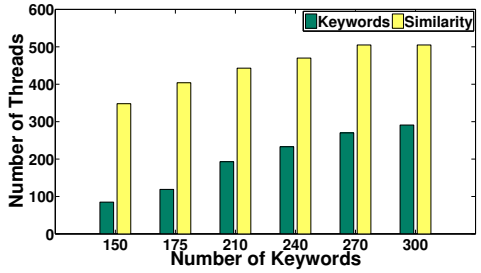


Figure 3: The robustness of the similarity approach to the initial keywords: number of relevant threads as a function of the number of keywords for OffensiveCommunity .

The features improves classification for all approaches. We briefly discusssed features in our classification section. We conduced experiments with and without these features for all four algorithms and we show the results in Figure 4 for OffensiveCommunity . Including the structural features in our classification improves the accuracy for all approaches (on average by 2.3%). The greatest beneficiary is the Bag-of-Words method whose accuracy improves by roughly 6%.

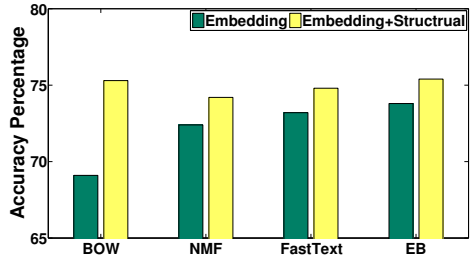


Figure 4: Classification accuracy for two different features sets in 10-fold cross validation in OffensiveCommunity .

EB outperforms the competition. Table 2 summarizes the comparison between the baseline methods and our Embedding Based approach (EB) for our three forums. EB consistently outperform other baseline method with at least 1.4 percentage point in accuracy and 0.7 percentage point in F1 score. Note that methods BOW and NMF did not assign any instances to the minority classes correctly, therefore the value of F1 score in Table 2 is reported as NA. Even though the EB outperforms other state-of-the-art methods, in the future work we are looking for possible approaches to involve the class labels in the embedding representation to improve the performance to the classification task.

Datasets	Metrics	BOW	NMF	FastText	EB
OffensComm.	Accuracy	75.3	74.2	74.8	75.4
	F1 Score	NA	NA	72.6	74.5
HackThisSite	Accuracy	66.6	72.4	69.7	74.6
	F1 Score	NA	70	65.7	72
EthicalHackers	Accuracy	59.9	58.2	59.9	61.1
	F1 Score	NA	57.2	58.9	59.5

Table 2: Classification: the performance of the four methods in classifying threads in 10-fold cross validation.

5 CONCLUSION AND FUTURE WORK

There is a wealth of information in security forums, but still, the analysis of security forums is in its infancy, despite several promising recent works. We propose a novel approaches to identify and classify IP addresses and threads of interest posted in security forums. In the future, we plan to extend our work by extracting other types of security information. Our future plans include:a) Considering class labels in embedding domain to improve the embedding performance in the classification. b) Developing a technique to automate keyword extraction from security forums to close the loop of identifying and classifying contents based on given keywords.

6 ACKNOWLEDGMENTS

I would like to thank to my PhD advisors, Professor Michalis Faloutsos for supporting me during these past four years.

REFERENCES

[1] Wenyuan Dai et al. 2007. Boosting for Transfer Learning *(ICML '07)*. New York, NY, USA, 193–200.
[2] Hal Daume III. 2007. Frustratingly Easy Domain Adaptation *(ACL '07)*.
[3] R. Frank et al. 2016. Location, Location, Location: Mapping Potential Canadian Targets in Online Hacker Discussion Forums *(EISIC '16)*.
[4] Joobin Gharibshah et al. 2017. InferIP: Extracting actionable information from security discussion forums *(ASONAM '17)*.
[5] Joobin Gharibshah et al. 2018. Mining actionable information from security forums: the case of malicious IP addresses. *CoRR* abs/1804.04800 (2018).
[6] Joobin Gharibshah et al. 2018. RIPEx: Extracting Malicious IP Addresses from Security Forums Using Cross-Forum Learning. In *PAKDD 2018*. Cham, 517–529.
[7] Thomas J Holt et al. 2012. Examining the social networks of malware writers and hackers. 6, 1 (2012), 891–903.
[8] Peng Jin et al. 2016. Bag-of-embeddings for text classification. *IJCAI International Joint Conference on Artificial Intelligence* 2016-Janua (2016), 2824–2830.
[9] Armand Joulin et al. 2017. Bag of Tricks for Efficient Text Classification. In *ACL 2017*. 427–431.
[10] Quoc Le et al. 2014. Distributed Representations of Sentences and Documents *(ICML'14)*. II–1188–II–1196.
[11] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (oct 1999), 788.
[12] Tomas Mikolov et al. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[13] Marti Motoyama et al. 2011. An Analysis of Underground Forums *(IMC '11)*. New York, NY, USA, 71–80.
[14] S. J. Pan and Q. Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.
[15] Rebecca S. Portnoff et al. 2017. Tools for Automated Analysis of Cybercriminal Markets *(WWW '17)*. 10.
[16] S. Samtani et al. 2015. Exploring hacker assets in underground forums *(ISI '15)*.
[17] Dinghan Shen et al. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *ACL 2018*.
[18] Nazgol Tavabi et al. 2018. DarkEmbed: Exploit Prediction with Neural Language Models. In *IAAI2018*.
[19] Guoyin Wang et al. 2018. Joint Embedding of Words and Labels for Text Classification. In *ACL 2018*. 2321–2331.
[20] Karl Weiss et al. 2016. A survey of transfer learning. *Journal of Big Data* 3, 1 (28 May 2016), 9.
[21] Xiong Zhang et al. 2015. The classification of hackers by knowledge exchange behaviors. *Info. Systems Frontiers* 17, 6 (2015).

Dynamics of Misinformation Cascades

Jasser Jasser
University of Central Florida
Orlando, Florida
Supervised By: Dr. Ivan Garibay
jasser.jasser@ucf.edu

ABSTRACT

The study of information-sharing cascades has been a constant endeavor since the emergence of social networks. Internet memes which mostly consist of catchphrases, viral images, or small videos shared over the social network are notorious for attracting the users' attention and spreading through the web in a fast fashion. Misinformation propagators latch their message to a meme to maximize the influence and spreading of the false news. As a result, the diffusion of misleading content has become a force to be reckoned with in the field of information warfare, as foreign actors seek to change opinions, manipulate ideologies, and create conflicts. In this study, we analyze the rapid dissemination of misinformation, aka, misinformation cascades, focusing on cascade temporal behavior and multi-cascade influence relationships. Twitter data used in this study contains only information associated with the Russian Internet Agency (IRA) and the Iranian Cyber Army (ICA). Our study focuses on analyzing temporal patterns of information dynamics created by these foreign actors for the sole purpose of spreading misinformation. We explore dividing temporal cascades into phases, where each phase differs from the previous regarding the number and characteristics of the information bursts. For this preliminary study, we are focusing on the #Trump and #USA hashtags used by the ICA. By studying the dynamics behind each phase, the forces behind the transition from one phase to another, and the influence relationships between cascades and their phases, we expect to shed some light on the timely subject of how to identify and protect society from information manipulation campaigns.

CCS CONCEPTS

• **Information systems → Social networks**; • **Networks → World Wide Web (network structure)**; **Online social networks**; • **Human-centered computing → Social networks**.

KEYWORDS

Misinformation Cascades, Cascade Phases, Misinformation Propagators, Cascade Influence

ACM Reference Format:
Jasser Jasser. 2019. Dynamics of Misinformation Cascades. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3314194

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314194

1 STATE-OF-THE-ART

The literature has addressed the issue of misinformation within multiple frontiers. In the political spectrum, researchers studied the effect of 'fake news' in polarizing opinions regarding a variety of issues such as healthcare reform [3] [9] by misleading the readers into believing in information that is factually incorrect. With the rise of social networks and social media, sharing political rumors over the web lead to harmful consequences of manipulating opinions [8]. Readers' Ideological biases pushed towards approving the false-shared messages and help in spreading it [4]. Researchers argue that readers usually are more interested in fake news where humans curiosity toward novel information is the main factor behind misinformation spreading, and false information was found to be more novel than true ones [14]. Readers' attitude towards political issues continues to shape even after discrediting the false message [11].

Information-sharing cascades are the primary mechanism by which false contents spread over the web and influence its audience. Researchers studied the cascade phenomenon; its temporal patterns, sharers' characteristics, and network structure. Rumor cascades tends to become more popular in social networks than truthful or factual ones [6], and are more likely to change over time where each burst in the cascade is dominated by different variants of rumors [7]. In general, "recurrence is widespread in the temporal dynamics of large cascades" by temporal, demographic, network, and different-copies features [5]. Recent research also tackled the message and its source to determine the mutability of the diffusing information [10]. However, these researches focused on studying the bursts within the cascades, its recurrence, characteristics, and the topic change in the recurred bursts. In our approach, we are attacking the issue from a different direction. This study focuses on dividing the cascade into phases where we investigate the dynamics in each phase and the dynamics change in the transition to the next phase.

Researchers also studied early detection of promoted campaigns [13], and the prediction of the popularity of user generated content [1] and its evolution [2]. Our study would lay the groundwork for developing an intelligent model that combines the characteristics of the hashtag cascade phases with the activities and influences of the users propagating the false messages to predict the temporal direction of the cascade. Early detection of a new phase in the misinformation campaign would help in determining its severity, the direction this hashtag will be used in propagating misinformation, and the proper way to countering it.

2 PROPOSED APPROACH

Our investigation is three-fold. First, investigating the characteristics and recurrence model of a hashtag cascade that was used by misinformation networks facilitated by foreign actors. The cascade characteristics will shed light on the temporal patterns, the propagator's aspects, and their network structure. Secondly, a closer look at the most active propagators within these cascades and how their influence affects the other propagators. Finally, studying the correlation between two hashtags, within the same misinformation network and outside networks, that share the same features, and whether one hashtag cascade played a role in influencing the other one into transitioning to a new phase.

2.1 Characteristics and Recurrence Model

For the misinformation cascade traits, an investigation of the temporal patterns of the hashtag cascade will be conducted by dividing the cascade timeline into phases. So far, from visualizing the phases, we categorized them into three categories

- Low activity phase: where no or limited bursts occurred during that period.
- Bursty phase: consists of an increase in amplitude and amount of bursts during that phase.
- Mixed-phase: where the temporal pattern is a mix between the two previously described phases.

Since this is a work in progress, currently we do not have a final operator to determine if a cascade is one these phases. We are currently exploring various alternatives on how to most appropriately operationalize these ideas.

2.2 Propagators activity and influence

We analyze the activity and influence of users who are propagating misinformation. We will refer to these users as propagators. We are looking to identify the major players that helped shape the burstiness of the cascade. So far, we identified four types of propagators; 1. Active and influential, 2. Active and trivial, 3. Inactive and influential, and finally 4. Inactive and trivial. The propagators' activities and influence will be monitored within the phases we defined in the previous section. This analysis seeks to determine the propagators rank within the cascade, and would provide an insight on how the propagators' activities and influence helped in the transition and the shaping of the studied phase.

In this section we are still to define a model to combine the propagators' activities and influences with the cascade phases to determine the criticality of the propagators in shaping the studied phase.

2.3 Do cascades influence other cascades?

Finally, we are looking to find whether there is a correlation or not between the sudden burstiness in one cascade and burstiness in other cascades that share a similar message. For this approach, we are looking into different hashtag cascades that share the same message and the effect of one cascade over the others. We are going to investigate any influence that might occur during the bursty phase of the cascade. A research question would be the following:

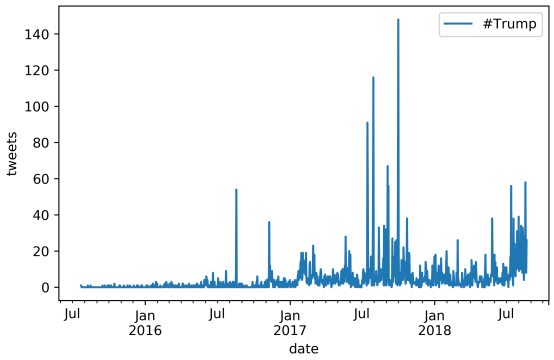


Figure 1: The cascade of the hashtag #Trump by the ICA agents

- Does a sudden burstiness in a hashtag cascade influence the sudden burstiness of another hashtag cascade, if they share the same characteristics?

3 METHODOLOGY & PRELIMINARY RESULTS

3.1 Dataset

Recently, Twitter disclosed information regarding accounts and related content associated with possible information operation that was discovered on the service since 2016 [12], enabling further research on information operations on Twitter. The data contains contents and accounts associated with the Russian Internet Agency (IRA) and the Iranian Cyber Army (ICA). The dataset comprises of 3,841 IRA accounts originating in Russia, and 770 ICA accounts from Iran with more than 10 million Tweets. This dataset will be used to analyze the hashtag cascades created by the both IRA and ICA users.

3.2 Cascade Phases

To determine the phases of the cascade, we used an offline change point detection algorithm to predict the points where the change in the cascade happens. In this example, we are studying the use of #Trump hashtag by the ICA agents. Fig.1 shows the whole cascade, while Fig.2 shows the cascade after determining the phases. Phase 1 is a mixed phase where there is a couple of bursts almost to the end of it. Phase 2 is a bursty one since the dynamics of the cascade changed drastically within that period to generate the most significant bursts. Finally, phase 3 is another bursty cascade but not as significant as phase 2 and is considered the beginning of a newly formed phase that will continue in the future. Notice that phase 2 started on January 2017, at the time of the inauguration, which leads us to believe that the ICA agents showed more interest in this hashtag where they used it for their misinformation campaigns.

3.3 Major Propagators

To determine the significant propagators within the cascade, we took into consideration the activity and influence of the propagators within the determined phases. Fig.3 shows the most active while Fig.4 shows the most influential users within the cascade.

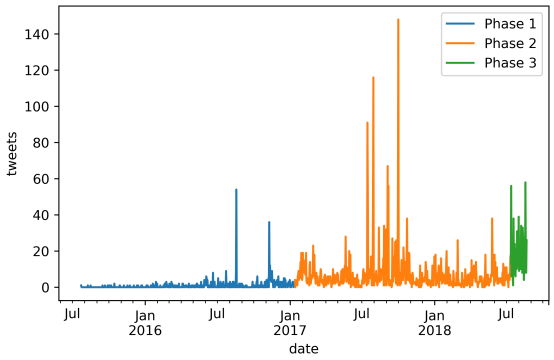


Figure 2: The #Trump cascade divided into phases

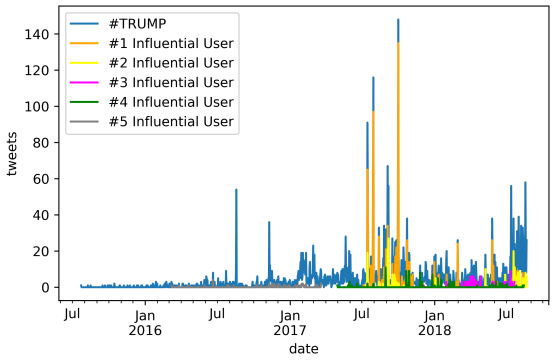


Figure 4: Most influential propagators within the #Trump cascade

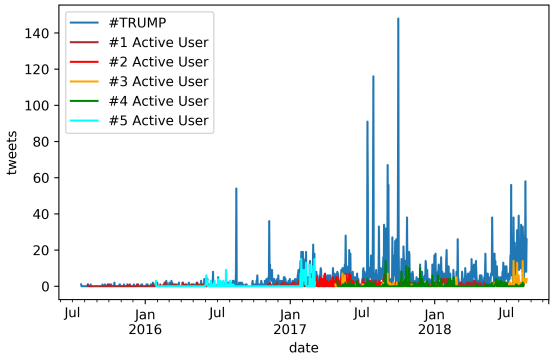


Figure 3: Most active propagators within the #Trump cascade

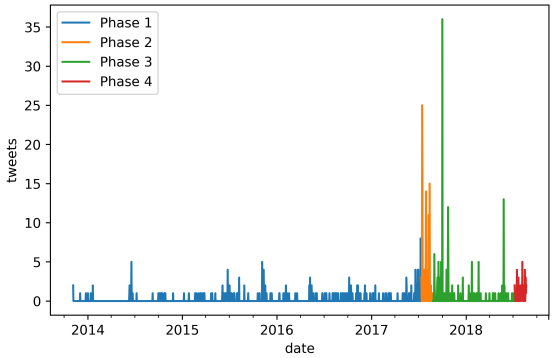


Figure 5: #USA cascade divided into phases

By influential, we mean any propagator that has a high number of reshares of its original message. The most and second most active users (Brown and Red lines in Fig.3) do not influence at all since nobody retweeted their messages. Thus, we cannot consider them as major players within the cascade. The third most active user is the most influential user (orange line in both figures), being somehow active within the second phase and most active in the third phase while its influence falls mostly within the second phase, making this user the dominant player of this cascade. Notice that the most influential user entered the cascade at the very beginning of phase 2 as shown in Fig. 3, at the time where this hashtag became interested in the ICA network. Finding a relationship between the significant players' activities and influences, and the change in the temporal pattern of the cascade is still a work in progress.

3.4 Cascade influence on other cascades

We monitored another hashtag (#USA) that was also used by the ICA agents. We also divided this cascade into phases as shown in Fig.5 We noticed that the two bursty phases (Phase 2 and 3) happened at the same time as the bursty phase of the #Trump cascade. We want to study whether the #Trump cascade had an influence on the #USA cascade within that period. In Figure 6, it shows that bursts in both cascades happened during the same period, but we need to make sure that the transitioning to the significant phases in the #USA cascade happened because of the critical phase that belongs to the #Trump cascade.

We are also interested in investigating the influence between both the IRA and the ICA networks. Whether IRA actors influenced ICA actors to act or vice versa. Such an investigation would provide a clear understanding of the change that incurred on a cascade where the influence did not happen from the same network. It also provides a closer look at the collaboration between the two networks, and how misinformation campaigns were developed and executed.

4 CONCLUSION AND FUTURE WORK

In this study, we are following a new approach to analyzing the temporal patterns, characteristics, network structure, and influence of misinformation cascade by dividing it into phases. Within each phase, propagators' activities and influences are analyzed to study the effect of these propagators on transitioning the cascade into different phases. We also explored the idea of finding whether a cascade can influence another cascade within the same network into burstiness. This work is far from over; there are still points

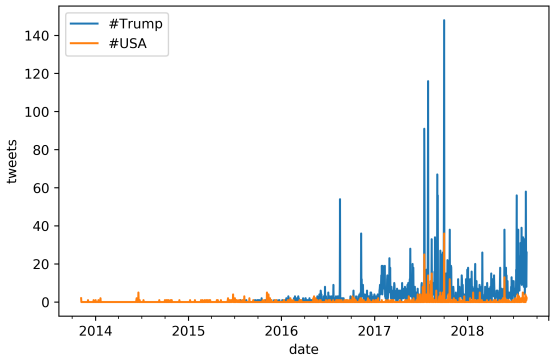


Figure 6: #Trump cascade and #USA cascade

that require an in-depth analysis to understand the full picture of the dynamics behind the formation of the cascades.

Our look into the future of this project is to be able to predict the next phase's temporal pattern. That will help in deciding whether a misinformation campaign using a specific hashtag to spread a false message should be given more attention or not. The combination of previous phases characteristics and temporal patterns with the propagators' activities and influence should provide us an insight into creating an intelligent model that would accomplish predicting the cascade future.

REFERENCES

[1] Khushnood Abbas, Mingsheng Shang, Alireza Abbasi, Xin Luo, Jian Jun Xu, and Yu Xia Zhang. 2018. Popularity and Novelty Dynamics in Evolving Networks. *Scientific Reports* 8, 1 (2018). https://doi.org/10.1038/s41598-018-24456-2

[2] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A peek into the future. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*. 607. https://doi.org/10.1145/2433396.2433473

[3] Adam J. Berinsky. 2017. Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science* (2017). https://doi.org/10.1017/S0007123415000186

[4] Michael A. Cacciatore, Sara K. Yeo, Dietram A. Scheufele, Michael A. Xenos, Doo Hun Choi, Dominique Brossard, Amy B. Becker, and Elizabeth A. Corley. 2014. Misperceptions in polarized politics: The role of knowledge, religiosity, and media. *PS - Political Science and Politics* (2014). https://doi.org/10.1017/S1049096514000791

[5] Justin Cheng, Lada A Adamic, Jon Kleinberg, and Jure Leskovec. 2016. Do Cascades Recur? (2016). https://doi.org/10.1145/2872427.2882993 arXiv:1602.01107

[6] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. https://doi.org/10.1073/pnas.1517441113 arXiv:1509.00189

[7] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades. *Icwsm* (2014).

[8] R. Kelly Garrett. 2011. Troubling Consequences of Online Political Rumoring. *Human Communication Research* (2011). https://doi.org/10.1111/j.1468-2958.2010.01401.x

[9] Brendan Nyhan. 2010. Why the "death Panel" Myth wouldn't die: Misinformation in the health care reform debate. *Forum* (2010). https://doi.org/10.2202/1540-8884.1354

[10] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior* (2018). https://doi.org/10.1016/j.chb.2018.02.008

[11] Emily Thorson. 2016. Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication* (2016). https://doi.org/10.1080/10584609.2015.1102187

[12] Twitter. 2018. Elections integrity. https://about.twitter.com/en_us/values/elections-integrity.html#2016-us-elections

[13] Onur Varol, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2017. Early detection of promoted campaigns on social media. *EPJ Data Science* 6, 1 (2017). https://doi.org/10.1140/epjds/s13688-017-0111-y arXiv:1703.07518

[14] S Vosoughi, D Roy, and S Aral. 2018. *the Spread of True and False News Online*. Technical Report 6380. 1146–1151 pages. http://science.sciencemag.org/

ClickGraph: Web Page Embedding using Clickstream Data for Multitask Learning

Porter Jenkins
Pennsylvania State University
prj3@psu.edu

ABSTRACT

The rise of big data frameworks has given website administrators the ability to track user clickstream data with more detail than ever before. These clickstreams can represent the user's intent and purpose in visiting the site. While existing work has explored methods for predicting future user actions, these methods are limited focus solely on one task at a time, ignore graph structure inherent in clickstreams, or model the conversion of the entire clickstream session, ignoring complexities such as multiple conversions in a single session. In this work, we formulate the novel problem of simultaneously predicting future user actions given a user's clickstream history. We argue that clickstream data contains important signal for predicting future user action. To tackle this new problem, we propose a novel method called `ClickGraph`, a recurrent neural network that encodes the graph structure of user click trajectories in the learned representations of web pages. We conduct experiments on a real-world dataset and demonstrate that this multitask learning approach is effective at improving the prediction of form fill conversions over strong baselines. In particular, we demonstrate that the ClickGraph model is effective at reducing false positive rates, increasing F1 scores, and improving recall.

KEYWORDS

Clickstreams, Recurrent neural networks, Representation learning, Multitask learning

ACM Reference Format:
Porter Jenkins. 2019. ClickGraph: Web Page Embedding using Clickstream Data for Multitask Learning. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314198

1 INTRODUCTION

Website administrators are keenly interested in understanding the behavior of visitors, or users, of their site. Better understanding of website traffic typically leads to better site design, increased sales, or improvement of other business goals. A litany of research in recommendation systems [17][13][2] and click-through-rate prediction [8] [16] relies on understanding or statistically leveraging past user behavior. The goal of many of these algorithms is to predict future user actions. If a program can successfully predict future user action, it would then be able to make intervention, and presumably drive improvement in some business outcome.

One important data feature to consider when predicting user actions is the clickstream. In this paper, we refer to a clickstream as a sequence of web pages a user visits. These pages are connected by a set of links, in a network- or graph-like structure. In many contexts, this clickstream data captures important characteristics about the user, and may often contain signal relevant to predicting future action. For example, [3] hypothesizes visitors to an e-commerce site augment their knowledge about the offered products as they traverse the site. We adopt this view, and believe it can be used to predict a variety of future user actions.

We extend this avenue of research and study the relationship between clickstream behavior and form-fill conversion rate. A form-fill is an event where the user provides information for future marketing contact. Many companies maintain a page where traffic from paid or organic search is directed. This page is designed to increase interaction with potential customers and generate new leads [5].

To the best of our knowledge, no existing work seeks to jointly model multiple user actions in clickstream data. In this work, we propose the novel problem of simultaneously predicting multiple user actions (multi-task learning) given sequential browsing history (clickstreams). Simultaneous prediction of future action allows for a multi-dimensional understanding of future behavior, and better informs understanding of user intent and preferences. There are significant challenges associated with solving this problem.

(1) It is difficult to effectively model the clickstream sequence and the relationships between different actions. For one, representing sequences of web pages in a meaningful way is a challenge. Many data scientists and researchers rely on naive representations of clickstream data such as discrete counts of site visits, number of pages clicked, session duration, average page duration, etc... [3]. Additionally, we cannot simply treat the different actions as independent. One action likely informs another.
(2) If independence is not a valid assumption, what is the best way to jointly model user actions, and in turn make simultaneous predictions? This is a non-trivial task that has yet to be solved in the literature.

We solve the problems in (1) and (2) by proposing a new method, `ClickGraph`, that is an LSTM-based model. As a Recurrent Neural Network (RNN), `ClickGraph` naturally models the variable-length, sequential nature of clickstream data. The model simultaneously predicts the next page the user will visit, the time duration, and form fill conversion. In this multi-task learning approach, we model the joint probability of actions by factoring conditional likelihoods in a way that mirrors the user's decision process. Additionally, we

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314198

incorporate graph structure into the loss of the next page prediction task, which informs other tasks of interest, such as form fill conversion. Finally, we use an embedding layer to learn representations of web pages to avoid manual, and often ineffective extraction of features.

We validate our proposed method against a variety of strong baselines and demonstrate that it is more effective at predicting form fill conversion. Conversion rates in a targeted marketing setting are notoriously low, where the average global conversion rate is approximately 2-3%, and the average cost (per dollar spent) to acquire a new customer is 92 cents [3]. Therefore, even marginal improvements in accuracy, or decreases in false negatives would have a dramatic impact on revenue. We show that, in addition to improving accuracy and AUC, `ClickGraph` is highly effective at decreasing false negative (FN) classifications and improving recall, two key metrics used in assessing the prediction of rare events.

In summary, the main contributions of our paper are as follows:

- We define a new problem: simultaneously predicting future user actions from sequential, clickstream data. Simultaneous prediction improves single tasks of high interest, such as form fill conversion prediction
- We solve this problem by proposing a new method, `ClickGraph`, that incorporates the graph structure of websites via a multi-task loss function. The model learns representations of web pages that encode this network information, and shares statistical strength across multiple user actions. These representations can be used by a machine learning model to better understand user intent.

2 RELATED WORK

Our problem touches on two separate literature streams, that both fall under of the umbrella of digital marketing: click-through rate prediction and clickstream models.

2.1 Click-through Rate Prediction Models

We discuss two related areas: sponsored search and content advertising [4]. In sponsored search, the user enters a query into a search engine and is served both organic search results, and relevant ads. In content advertising, ads are displayed alongside original content of a website, where the ads can either be generic or targeted to the user [4]. Rather than predict ad click-through rates, we address the problem of predicting future actions a user will take within a site.

Recent work focuses on learning representations of users using neural networks in an Embedding&MLP approach [8] [13] [8] for predicting digital ads clicks. Additionally, some approaches model sequential data [16]

2.2 Clickstream Models

Historically, there are four common approaches to model clickstream data of web site visitors: sequential pattern analysis, association rules, clustering, and Markov models [7]. All of these methods model a conversion event (click, buy, form-fill, etc.) for an entire sequence of clickstream data for any given session. The granular temporal information of conversion, or even decoding multiple conversions in a single session, is unaccounted for in these methods.

To the best of our knowledge, no work to date tries to solve the problem of simultaneously predicting multiple user actions from sequential clickstream data.

3 PROBLEM DEFINITION

In what follows, we formally define our problem in terms of inputs and outputs. The input to our problem consists of three major components, a website graph, user clickstreams, and user actions.

3.1 Inputs

Definition 3.1 (Website graph). We treat the website of interest as a graph, $\mathcal{G}$, with nodes, or webpages, $\mathcal{P} = \{p_1, p_2, ..., p_m\}$ and edges, $\mathcal{E} = \{e_1, e_2, ..., e_q\}$. A given web page, p_i, is associated with a set of adjacent pages, or adjacency list, $\mathcal{A}_i = \{p_1, p_2, ..., p_k\}$.

Definition 3.2 (User actions). Additionally, we observe a set of n users, $\mathcal{U} = \{u_1, u_2, ..., u_n\}$, and user actions $\mathcal{Y} = \{Y_{1,1}, ..., Y_{n,t}\}$. For each of the n users, at each timestamp, t, we observe a tuple of actions $Y_{it} = \langle y_{it}, d_{it}, p_{i,t+1} \rangle$, where:

- y_{it}: (*Form fill*) Indicator denoting whether the user requests marketing information (fills out a form). Where, $y_{it} \in \{0, 1\}$
- d_{it}: (*Duration*) Time (in seconds) user u_i spends on page p_m. Where $d_{it} \in \mathbb{R}^+$
- $p_{i,t+1}$: (*Next page*) The future page user u_i will visit at time $t + 1$. Where $p_{i,t+1} \in \mathcal{A}_i$

Definition 3.3 (User clickstream). Finally, we observe a set of n clickstreams, $\mathcal{S} = \{S_1, S_2, ..., S_n\}$, where S_i is a sequence of web pages and actions at time t,

3.2 Outputs

We propose a model that takes input as described and produces probability estimates of multiple tasks. Our primary goal is to predict future user actions. Thus, the first output is a tuple of probabilities over future user actions, $\hat{\mathcal{Y}} = \{\hat{Y}_{1,t+1}, ..., \hat{Y}_{n,t+f}\}$, where $\hat{Y}_{n,t+f} = \langle P(y_{i,t+f}), P(d_{i,t+f}), P(p_{i,t+f+1}) \rangle$, and f denotes the number of future time periods predicted. As will be shown in section 4, we do not model these actions independently, but do so jointly mirroring the user's decision process.

4 PROPOSED METHOD

In this section, we provide an intuitive overview of our model, and give a very brief outline of Recurrent Neural Networks (RNN's) and Long Short-term Memory (LSTM) cells. We then provide a more in-depth discussion of our proposed method, `ClickGraph`, it's architecture, and define its loss function.

4.1 ClickGraph Overview

The `ClickGraph` model is a recurrent neural network that takes sequences of web pages as input and outputs probabilities of the three potential actions described in section 3. At each time stamp, to compute these probabilities we model the joint likelihood of all three possible actions, $P(y_{it}, d_{it}, p_{i,t+1})$. We factor this joint probability to reflect the decision process of the user. Specifically, the next page the user will visit is dependent on how long they spend on the page, and whether or not they fill a form. Additionally,

page duration is also conditional on form fill conversion. Finally, form fill conversion is dependent on the hidden state learned by the neural network. We provide a more rigorous definition in 4.3.4.

4.2 Recurrent Neural Networks and LSTM

Recurrent Neural networks are powerful deep learning models that process sequential data. Modern networks have proven effective in many tasks because they share weights across different parts of the model, and across time stamps [6]. This sharing enables better generalization and makes them more robust to examples appearing at different time stamps.

In practice, some of the most effective RNN's are those that utilize Long Short-Term Memory cells [10]. The key idea behind LSTM models is their ability to learn when to forget older states, which in turn help them alleviate the effect of gradient vanishing. LSTM's are also able to effectively manage the trade-off between long- and short-term dependencies [6].

4.3 ClickGraph Architecture

We now define the architecture of our proposed model, `ClickGraph`.

4.3.1 Input Layer. A single input to our model is a given user's clickstream, $S_i = ((p_{i,1}, Y_{i,1}), ..., (p_{i,t}, Y_{i,t}))$. At each time, t, we observe the page the user traverses to, p_i, and the set of actions taken, $Y_{i,t}$. Our goal is to predict these actions, $Y_{i,t}$, from p_i on unseen data. At each time, t, we input a one-hot representation of $\mathbf{p}_{i,t} \in \mathbb{R}^{|\mathcal{P}|}$. Note that the length of S_i varies with respect to users.

4.3.2 Webpage Embedding Layer. The input $\mathbf{p}_{i,t}$ is extremely sparse and is a poor representation of relationships between pages on the website. Pages represented as one-hot vectors all have the same euclidean distance ($\sqrt{2}$) from each other in $\mathbb{R}^{|\mathcal{P}|}$ [6]. Therefore we seek to learn a representation that encodes similarity between pages drawing on ideas from [9][16], by learning a lower dimensional page representation, $\mathbf{e}_{i,t} \in \mathbb{R}^{\delta}$, via a linear mapping:

$$\mathbf{e}_{i,t} = \mathbf{W}_{embed}\mathbf{p}_{i,t} \tag{1}$$

where $\mathbf{W}_{embed} \in \mathbb{R}^{\delta \times |\mathcal{P}|}$ is a matrix of learnable parameters.

In section 4.3.5 we define our loss function, which contains a term for next page prediction. This term incorporates the adjacency matrix of the graph in the softmax denominator. Because the weights, $\mathbf{W}_{embed}$, are updated through propagation of errors, optimization steps that minimize the loss function in turn encode graph information into the embedding vectors, $\mathbf{e}_{i,t}$.

4.3.3 LSTM Layer. The web page embedding, $\mathbf{e}_{i,t}$, is fed into an LSTM cell, which produces a hidden state at time, t. For brevity, we refer the reader to [10][6][11] and omit the LSTM definition here.

4.3.4 Multi-task Prediction Layer. Once we have obtained the hidden state vector $\mathbf{h}_t$ from the LSTM cell, we feed it into the output layer for predicting $y_{i,t}$, d_{it}, and $p_{i,t+1}$. We model the joint likelihood of the three user actions using the chain rule of probability:

$$\begin{aligned} &P(y_{i,t}, d_{it}, p_{i,t+1}) = \\ &P(y_{i,t}|h_t)P(d_{it}|y_{i,t}, h_t)P(p_{i,t+1}|d_{it}, y_{i,t}, h_t) \end{aligned} \tag{2}$$

We then encode these conditional probabilities into the architecture of the network.

Task 1: Form Fill We obtain the predicted form fill probability from equation 3

$$P(y_{i,t}|h_t) = \sigma(\mathbf{W}_{form}\mathbf{h}_t + \mathbf{b}_{form}) \tag{3}$$

Where $\sigma(\cdot)$ is the sigmoid function [6]. We then condition on this estimate of $\hat{y}_{i,t}$, to predict $P(d_{it})$ and $P(p_{i,t+1})$ as in (2).

Task 2: Page Duration

To better predict our other tasks, we first embed the predicted value for form fill into a distributed representation, $\boldsymbol{\gamma}_{it} \in \mathbb{R}^{p \times 1}$. This distributed representation eliminates the presence of a binary component of the feature vector when predicting $d_{i,t}$ and $p_{i,t+1}$. We let the dimensions of $\boldsymbol{\gamma}_{it}$ be $p = \left\lfloor \frac{\delta}{2} \right\rfloor$

$$\boldsymbol{\gamma}_{it} = \Omega\hat{\boldsymbol{y}}_{it} \tag{4}$$

where $\Omega \in \mathbb{R}^{p \times 2}$ is a matrix of learnable parameters, and $\hat{\boldsymbol{y}}_{it} \in \mathbb{R}^{2 \times 1}$ is the vector of predicted probabilities for form fill conversion.

This embedded vector is concatenated with the hidden state variable, $\mathbf{h}_t$, to produce a feature vector for the regression of predicted page duration:

$$\boldsymbol{\beta}_t = \begin{bmatrix} \boldsymbol{\gamma}_t \\ \mathbf{h}_t \end{bmatrix} \tag{5}$$

We assume a gaussian distribution with mean, μ and unit variance for $P(d_{it})$,

$$P(d_{it}|y_{i,t}, \mathbf{h}_t) \sim \mathcal{N}(\mu, 1) \tag{6}$$

where, $\mu = \mathbf{w}_{duration}^{\top}\boldsymbol{\beta}_t + \mathbf{b}_{duration}$

We derive our predicted values for $\hat{d}_{i,t}$ from the dot product of the weight and feature vectors for page duration:

$$\hat{d}_{i,t} = \mathbf{w}_{duration}^{\top}\boldsymbol{\beta}_t + \mathbf{b}_{duration} \tag{7}$$

Task 3: Next Web Page

Finally, we concatenate the normalized predicted value, $\hat{d}_{i,t}$ (scalar), with the embedded predicted values for $\hat{y}_{i,t}$ to obtain the feature vector, $\boldsymbol{\pi}_t$, for the next page prediction task.

$$\boldsymbol{\pi}_t = \begin{bmatrix} \boldsymbol{\gamma}_t \\ \hat{d}_{i,t} \end{bmatrix} \tag{8}$$

We take the softmax of the feature vector and weight matrix over the adjacency matrix with the softmax adjacency function:

$$\hat{p}_{i,t+1} = softmax_{adj}(\mathbf{W}_{page}\boldsymbol{\pi}_t + \mathbf{b}_{page}) \tag{9}$$

Where,

$$softmax_{adj}(\mathbf{x}_i) = \frac{e^{x_i}}{\sum_{k \in \mathcal{A}_i} e^{x_k}} \tag{10}$$

and $\mathcal{A}_i$ is the adjacency list for page, p_i. This allows us to encode graphical structure into our sequential predictive model.

4.3.5 Loss Function. In the following section we define the loss function used to train ClickGraph. Our objective is to minimize the negative log likelihood in equation 11. Additionally, we insert the tuning parameter $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3]$ to control the contribution of each task to the overall loss. Using cross-entropy loss for $P(y_{i,t})$ and $P(p_{i,t+1})$, and mean squared error loss for $P(d_{it})$. We add an an L2 penalty over the weights for $p_{i,t+1}$, which is controlled by another hyperparameter, α. This additional term acts as a regularizer on the next page prediction task and is very effective at improving generalization performance in our empirical tests.

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{t=1}^{T}[\ \lambda_1\left(y_{it}\log P(y_{i,t}) + (1-y_{it})\log(1-P(y_{i,t}))\right) + \frac{\lambda_2}{2}(\hat{d}_{it} - d_{i,t})^2 + \lambda_3 \log P(p_{i,t+1})\ ] + \frac{\alpha}{2}\|\mathbf{W}_{page}\|_2^2 \tag{11}$$

5 EXPERIMENTS

In the following section we describe the clickstream data set and experimental settings used [14] [1]. Additionally, we empirically evaluate our model against a variety of baselines, comparing against multiple metrics of performance. Finally, we provide a qualitative analysis of the embedding vectors learned by ClickGraph.

5.1 Experimental Settings

5.1.1 Dataset Description. We collect data from a large software-as-a-service company whose primary product is a data analytics platform. A portion of their site is targeted at prospective customers and is comprised of content describing the function and use of their software. We observe incoming users to this portion of the site, their clickstream trajectory, how long they spend on each page, and whether or not they click a form to provide contact information for marketing purposes. For our experiments, we structure the data in two ways. First, we maintain a tabular array of the dataset, where each transaction (row) is treated as independent from all other transactions. We rely on this data structure to feed data into baselines that ignore sequential information, or treat each observed transaction as *iid*. Additionally, we maintain a sequential data structure, where each sequence is indexed by a visit ID. The sequential data is used to easily collect sequence-based mini batches for the sequence models used in our experiments.

5.1.2 Prediction Task. We frame our prediction problem in the context of digital marketing, as discussed above. While our method can be generalized to learn about any future action facing a website user, we evaluate our model on the basis of form fill prediction, which is a binary prediction task. Form fill prediction is important because it provides insight into the user's intent.

We split our data into three sets: training (80%), validation (10%), and test (10%). For models that treat the data as *iid*, we split our tabular data structure along the rows of the design matrix. For models where sequential information is meaningful, we split our sequence data by randomly allocating *visit id*'s to the train, validation, or test set.

5.2 Quantitative Evaluation

5.2.1 Baselines. To assess the effectiveness of our model we compare to state-of-the-art baselines that vary both the sequential and graphical aspects of the data. This allows us to validate our hypothesis that incorporating graph structure into a sequential model to simultaneously predict multiple variables is effective for predicting user action. Specifically, we compare the following methods:

- **MajorityClass** A naive baseline that unconditionally predicts the majority class (e.g., $\hat{y} = 0$), regardless of sequence.
- **XGB (One Hot)** A simple classifier with no sequence or graph information. We use an XGBoost algorithm with a naive, one-hot encoding of clickstreams. Each webpage is a one-hot vector of dimension $|\mathcal{P}|$. We treat each sample of our feature matrix as *iid*.
- **XGB (Graph)** Classifier with graph information, but no sequential information. We train an XGBoost algorithm with embedded vectors of clickstreams. We rely on the DeepWalk algorithm [15] to obtain graph-based embedding vectors of webpages. We treat each sample of our feature matrix as *iid*.
- **LSTM (No Graph)** A deep learning classifer that incorporates sequential information, but graphical information. This model is a simpler form of ClickGraph, but only predicts $y_{i,t}$, and does not incorporate the graph structure of the website.
- **ClickGraph** The proposed classifier that includes both sequential and graphical information.

5.2.2 Evaluation Metrics. Given that the prediction task is binary, we utilize standard binary classification metrics.

In our data, observing a form fill event is rare; approximately 2.3% of all samples have positives labels. In the context of our problem, a key interest is in identifying true form fill events. Because a trade-off exists between false positives and false negatives [12], we weight the cost of a FN as larger than that of a FP. As such, we primarily are interested metrics such as AUC, recall, and F1. A desirable model would have the property of successfully lowering FN's, without sacrificing overall effectiveness.

5.2.3 Results. We report the results form our quantitative evaluation in table 1. Overall, we have the following three observations.

Modeling the clickstream sequence improves performance The results in table 1 demonstrate that, in general, discarding the sequential structure of the data is harmful to model performance. Clearly, the sequence models based on recurrent neural networks (LSTM (no graph) and ClickGraph) outperform the XGBoost models in terms of AUC and accuracy. AUC tends to be a reliable overall measure as it summarizes the trade-offs between precision and recall. In the case of ClickGraph, the hidden layers are able to learn a representation of this intent in order to better identify users who will click "form fill."

Modeling the graph structure improves performance In our experiments we test two different variants of the classic XGBoost algorithm, as well as two variants on an LSTM. In both cases, graph information is incorporated in one, and not the other. Table 1 shows that for the XGBoost models, XGBoost (graph) outperform XGBoost (Naive) in terms of AUC, and accuracy. Additionally, we can make a similar observation between the LSTM-based models LSTM (no

Model	AUC	F1	Accuracy	Precision	Recall	FN rate
MajorityClass	0.5	0.0	0.9768	0.0	0.0	1.0
XGB (naive)	0.7989	0.2022	0.9725	0.3103	0.15	0.85
XGB (graph)	0.8592	0.0187	0.9756	0.1429	0.01	0.99
LSTM (no graph)	0.9388	0.0082	0.9812	**1.0**	0.0041	0.9959
ClickGraph	**0.9493**	**0.2913**	**0.9829**	0.6716	**0.1860**	**0.8140**

Table 1: Test set evaluation metrics for various classifiers. We observe that the proposed method, ClickGraph, outperforms all baselines along key metrics of interest: AUC, F1, Recall, and FN rate. For all relevant metrics, we use a classification threshold of .5

graph) and ClickGraph. The proposed model outperforms the more naive LSTM along a host of metrics: all but precision. While we do see obvious tradeoffs between FP's and FN's when analyzing the graph-based and graph-free models, the overall differences in AUC gives us insight into the nature of these tradeoffs.

Simultaneously predicting different actions reduces FN's and improves recall Note that the two models that yield the best AUC are the two recurrent neural networks, LSTM (no graph) and ClickGraph. While the AUC and accuracy metrics between the two are somewhat comparable, ClickGraph dramatically outperforms the naive LSTM in terms of F1, recall, and FN rate. This result seems to suggest that the multi-task aspect of the model, wherein we can encode the graph structure of the website through next page prediction, aids in reducing false negatives.

Additionally, it is worth noting the high precision of LSTM (no graph) as compared to ClickGraph. The perfect precision for LSTM (no graph) occurs because it predicts only one TP and zero FP's. In other words, the model is extremely biased towards $\hat{y} = 0$. In fact, it is hardly different the MajorityClass model. Conversely, ClickGraph dramatically increases TP's and decreases FN's. While it does incur more FP's, the cost of this error is low in the context of our problem. Additionally, overall model performance is maintained, demonstrated by high AUC and accuracy. As mentioned above, these results suggest that ClickGraph is a highly desirable model in the context of predicting form fill conversion.

6 CONCLUSION

In this paper, we studied user web behavior in a digital marketing context. We defined the new problem of simultaneously predicting multiple user actions from sequential, clickstream data. This multi-task approach to predicting future user actions is effective at improving prediction performance of high interest tasks, such as such as form fill conversion. We proposed a novel method called ClickGraph to fully utilize sequential, and graph structure in clickstream data. This same method can easily be generalized to predicting a variety of other user actions. We evaluated our method against strong baselines on real-world clickstream data. Our experiments demonstrate that ClickGraph is very effective at reducing false positives, and improving recall. We hope to explore other data sets and other settings, in particular different actions sets, for future work.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] et al. Chantat Eksombatchai. 2018. Pixie: A System for Recommending 3+ Billion Items to 200+ Million Users in Real-Time. *WWW 2018: The 2018 Web Conference, April 23-27, 2018, Lyon, France* (2018), 10.

[3] Ram Kosuru Choudur Lakshminarayan and Meichun Hsu. 2016. Modeling Complex Clickstream Data by Stochastic Models: Theory and Methods. In *WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.* ACM, 7 Pages.

[4] Ewa Dominowska and Vanja Josifovski. 2008. First Workshop on Targeting and Ranking for Online Advertising. In *WWW'08 Companion, April 21–25, 2016 Beijing, China.* ACM, 2 Pages.

[5] Optimezely Editors. 2018. Landing Page Optimization. Retrieved November 3, 2018 from https://www.optimizely.com/optimization-glossary/landing-page-optimization

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org.

[7] Sule Gunduz and M. Tamer Ozsu. 2003. A Web Page Prediction Model Based on Click-stream Tree Representation of User Behavior. In *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM, 6 Pages.

[8] Chengru Song Ying Fan Han Zhu Xiao Ma Yanghui Yan Junqi Jin Han Li Guorui Zhou, Xiaoqiang Zhu and Kun Gai. 2018. Deep Interest Network for Click-through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM, 10 Pages.

[9] Rami Al-Rfou Haochen Chen, Bryan Perozzi and Steven Skiena. 2018. A Tutorial on Network Embeddings. *arXiv preprint arXiv:1808.02590v1* (2018).

[10] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-term Memory. *Neural Computation* (1997), 9(8):1735–1780.

[11] Jan Koutnik Bas R. Steunebrink Klaus Greff, Rupesh K. Srivastava and Jurgen Schmidhuber. 2015. LSTM: A Search Space Odyssey. *Transactions on Neural Networks and Learning Systems* (2015).

[12] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective.* The MIT Press, Cambridge, Massachusetts.

[13] Jay Adams Paul Covington and Emre Sargin. 8. Deep Neural Networks for YouTube Recommendations. In *RecSys'16 September 15-19, Boston, MA, USA.* ACM, 10 Pages.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[15] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14).* ACM, New York, NY, USA, 701–710. https://doi.org/10.1145/2623330.2623732

[16] Ruofei Zhang Zhongfei (Mark) Zhang Shuangfei Zhai, Keng-hao Chang. 2016. DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks. In *Proceedings of the 22th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM, 10 Pages.

[17] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence* (2009).

Finding Themselves between Home and Host Cultures

The Use of Facebook by International Students

L. Malone
Media/Communications
University of Western Australia, Perth, Western Australia, Australia, lin.malone@research.uwa.edu.au

ABSTRACT

Advancements in technology have enabled society to become increasingly globalised, both with regard to physical migration, as well as through the use of information and communication technologies (ICTs) to enable the maintenance of transnational ties. In particular, transient migration in the area of higher education has seen an increasing number of students migrate overseas for the purpose of their studies. However, research has shown that these international students are often disconnected from their host culture and society, with local-international friendships proving to be uncommon (Baldassar & McKenzie, 2016). Based on interviews with over 200 international students in Australia, Sawir et al. (2008) revealed that two-third of them had suffered or were suffering from loneliness, of which Sawir et al. identified three kinds -personal loneliness from loss of contact with families, social loneliness from the loss of networks, and cultural loneliness from the change in environment. This raises the question of how these students may be better supported, and the international student experience improved upon.

Social media has often been positioned as a tool through which users become connected and communities are formed. One of the most popular platforms of social media, Facebook, has become an established part of many lives in modern society. Media and culture have always been interconnected, however, the dominance of the Web in everyday lives means that the role that media plays in cross-cultural communication is more significant than ever and must be researched for a better understanding of this phenomenon. While current research has examined the issues relating to construction of online identities for communication within established social networks, new issues have emerged in relation to collapsed contexts and imagined audiences in today's globalised world, especially as multiple cultures are introduced onto the same platform as a result of migration or relocation. Insufficient research has been done into the influence of technology on transient migration and its potential to support cross-cultural communication. The question thus is how exactly social media may assist transient migrants overcome issues of isolation and loneliness, and provide them with support during their time abroad.

This study looks to address the issue of student isolation within host societies by examining how social media may provide spaces for support, self-expression and cross-cultural communication. Through a visual internet ethnographic study, it examines the profile pages of international students on Facebook to better understand their positions in these home and host societies. This research study is supplemented by semi-structured interviews for a thorough examination of international students' use of social media.

As we examine how the Web has developed and changed over the last 30 years, it must be acknowledged that this change cannot be solely attributed to technological advancements, but is also influenced by the actual human users of the Web who participate in it. Web users utilise online tools to produce content of their own, tailoring their online experiences accordingly. While advancements in technology has created a more globalised society, the globalised users within this society have had their own impact on technology. This research looks to create a more in-depth understanding of the ways in which social networking platforms are used by transient migrants to navigate transnational cultural settings. It is aimed at enabling deeper understanding of the complex inter-linkages between cultures, to engender new insights of transnational identities. This is essential to address the global nature of today's society and the role of social media platforms in the spaces they create for transient migrants.

CCS CONCEPTS

• Human-centered computing~Social media • Human-centered computing~Ethnographic studies

KEYWORDS

Facebook, Social Media, Migration, International Students, Hybridity, Online Spaces

ACM Reference format:

L. Malone. 2019. Finding Themselves between Home and Host Cultures: The Use of Facebook by International Students. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3314193

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314193

1 State of the Art

In order to examine role of social media in facilitating the international student experience, this paper begins by examining the concepts of migration and integration through Berry's (1997) theory of acculturation and their critiques. Phinney (1996) describes acculturation as the extent to which individuals acclimate to host cultures, while maintaining their home culture identities. Berry draws on this to suggest a quadric-modal model that focuses on two dimensions of acculturation; maintenance of home cultural identity and acclimatisation to the host culture. Based on these two dimensions, Berry (1997) describes four modes of acculturations: (1) assimilation - where the individual interacts with the host culture and rejects their native cultural identity; (2) integration - where the individual interacts with the host culture and still maintains their native cultural identity; (3) separation - where an individual avoids participation with their host culture and focuses on maintaining their native cultural identity; and finally, (4) marginalization - where the individual does not interact with both their native culture and their host culture. Of the four, Berry (1997) views the ideal state of acculturation to be integration.

Indeed, the use of information and communication technologies (ICTs) within migration processes have often been seen to have positive impacts on acculturation. An example of Berry's integration can be read through Chen and Hanasono's study on the uses of Facebook and Renren by Chinese international students in America (2016). In this study, the Chinese international students utilised Facebook as an information tool to keep their selves updated with local news and with social happenings in their host society. Their continued use of Renren however, served more social purposes and was essential in allowing them to communicate with their friends and family back home. Indeed, research on ICTs have revealed them to be key tools in allowing migrants to maintain relationships despite issues of distance and space (Baldassar and Merla, 2014), as evidenced in Chen and Hanasono's study. The integration process can thus be seen to be supported by the use of ICTs.

However, the use of ICTs by international students to communicate with friends and family may also have negative impacts on their acculturation. Kim et al., (2009) researches the uses of communication technologies by international students in South Korea. Their findings revealed that the international students' preoccupation with maintaining their established relationships back in their home culture, meant that they lacked incentive to form new relationships within their host culture. Their social relations were limited to their friends and family back home, which impacted their social adjustment within their new host society, resulting in separation as described within Berry's (1997) model.

While Berry's model may be utilised to describe some of the processes that are taking place, there are limitations to its framework. Particularly, Berry's model has been critiqued for the narrowness of its approach to the migration process. Ngo (2008) argues that basing an acculturation model on two dimensions overlooks the impact that other factors may have on the acculturation process, while Rudmin (2003) critiques the model's "one size fits all" approach, which categorises all migrants equally. Weinreich (2009) argues that the model reduces the complex nuanced acculturation process into a simplified choice of acceptance or rejection of host cultural identities. While acculturation may have once served as an ideal framework through which to examine migration, the nature of globalisation today alongside the rise of social media, demands a shift in the ways in which we think about migration.

An alternative perspective is offered by Gomes (2014, 2015), whose work examines the use of social media by international students in Australia. Gomes, Berry, Algouzal and Chang's work (2014) reveals that international students form social identities tied to their host nation while studying abroad, linking back to the findings of Sawir (2008) and Kashima and Loh (2006). These social identities focus on the students' position as international students, and allow for the formation of "a parallel society made up of fellow international students that has no clear connections to (multicultural) Australian society or culture" (Gomes et al., 2015, p.517), maintained through social media and face-to-face contact. The formation of these social networks mimics the 'bubble' hypothesis in studies of expatriate communities, where expats tend to live in a 'bubble' separated off from the host society (Fechter, 2007).

To examine this phenomenom through the Berry's framework, this international student parallel society could be described as marginalisation in Berry's model. However, as argued by Del Pilar and Udasco (2004), home and host societies have a dominant influence on identity formation, making it impossible for migrants to construct identities that are completely independent of either culture. The students still associate with their home cultural networks, demonstrating the insufficiency of the marginalisation category to describing the positions held by these international students. In fact, none of the categories in Berry's model are quite able to convey the hybridity of the positions occupied by these students.

Gomes' research (2015) is essential in applying the notion of the hybridity to a theory of identities. The home and the host cultures are viewed as being intrinsically intertwined, meaning that these international students hold complex diverse identities that are not linked singularly to either their home or host cultures, but rather, draw from both (Gomes, 2014). In doing so, Gomes' work highlights the insufficiency of Berry's model as a whole, but particularly with regard to approaching research on the use of social media by international students solely from an integration perspective. Through the international student identity-based networks that these students occupy, it can be seen that integration is often not the end goal of most international students (Montgomery and McDowell, 2009). A more ideal approach would be to perform in-depth investigations of the hybrid positions that international students occupy in their home and host societies, and to utilise these understandings to contextualise the students' uses of social media as hybrid spaces tailored to their own needs.

2 Proposed Approach

Studies into the use of social media by international students have been limited in their scope, primarily due to their focus on the use of social media as a communicative platform. However, this limits understandings of the ability and potential of the Web to provide multiple spaces for information, social interaction, cross-cultural communication and identity production, especially to migrants who lack these spaces in offline environments. There is a need for more in-depth understandings of how international students may utilise social media to assist them in navigating the cultural differences that exist between their home countries and their countries of study. This research project aims to fill this gap by examining the Facebook platform as a Third Space – a platform that exists between home and host culture. This research draws on Bhabha's conceptualisation of the Third Space (1994) as an in-between hybrid position. As culture is only constructed in relation to the "Other", the third space serves as "an ambivalent site where cultural meaning and representation have no 'primordial unity or fixity'" (Bhabha, 1994, p.37). This allows for the alternative ways of seeing to emerge, in contrast to dominant viewpoints.

Gomes, 2014, 2015) and Bhabha's work (1994) supports a shift away from thinking of migrant positions in terms of the limited dimensions of Berry's model, to address migration from a hybridised perspective instead. There needs to be a change in understanding migrant identities beyond wholesale acceptance or rejection of entire cultures, to focus instead on how migrant users navigate the complexities of various cultures in their own ways. Beck and Beck-Gernshiem's work on individualization in today's society (2002) is particularly relevant in highlighting how "[t]hose who live in this post-national, global society are constantly engaged in discarding old classifications and formulating new ones. The hybrid identities and cultures that ensue are precisely the individuality which then determines social integration. In this way, identity emerges through intersection and combination and thus through conflict with other identities" (p.26). Online identities are thus subject to the agency of the individual and their choices of cultural content as relevant to their identities. This is particularly essential to understanding the uses of social media by international students as their identities and their perceived positions in the society shape these uses. Approaching Facebook as a Third Space looks to take into account the agency of migrants in online environments, as they tailor online spaces according to their own usage.

To examine Facebook as a hybridised Third Space, it is necessary to examine both user activity and content produced on the site. As such, beyond asking questions about the use of social media by international students, this study employs a visual Internet ethnographic approach to perform a detailed analysis of the content produced through the participants' Facebook profiles. This approach is aimed toward examining how experiences are reproduced within online identities, as well as to incite understandings of cultural interactions and communities of practices online. This ethnographic study is supplemented with semi-structured interviews into user activity on Facebook to contextualise their experiences and allow for insights into how the students are able to tailor their social media experiences.

3 Methodology

The participants gave access to their Facebook profile pages by accepting the researcher as a "Friend" and data from their profile pages were analysed in a visual Internet ethnographic study. The practice of visual Internet ethnography examines "a whole new raft of cultural practices mediated by visual and Internet technologies, through which the meaning and the significance of visual documents are very often redefined" (Ardévol, 2012, p.6). Facebook has fixed tools for user communication and set layouts for users' profile pages; which allow for a study on the construction of online identities through utilising the tools made available. However, despite Facebook's predetermined layout, the practices which individuals engaged in to create their identities on the site are unique to each participant as individuals; the visual set-up being intrinsically tied to their constructed digital identities. Thus images are as important and meaningful as written text in supporting detailed ethnographic work (Pink, 2013), an observation that emphasises the choice of visual ethnography as an effective research methodology. This methodology enables the visual and critical examination of the participants' profiles with regard to the presented text and images working together, from the position of an observer.

This visual Internet ethnographic study was further supported by face-to-face semi-structured interviews, where the students were asked questions about their experiences abroad, and the role played by social media in these experiences. During these interviews, the participants were encouraged to access their social media pages to reinforce any points that they made. These interviews were necessary to engage with the users in a more in-depth manner beyond what a study of their user practices may have afforded, to allow for further insights into understanding the use of social media by international students.

4 Results

While data collection and analysis are currently still ongoing, this research project has examined 29 international students in Finland and 21 international students in Australia to date. The data collected so far reflects the importance of social media to international students during their time abroad. The initial findings reflect some of the key ways in which social media are utilised by these international students as highlighted within the semi-structured interviews conducted:

To connect with friends: "I think [Facebook] is the main way I communicate with [people back home]" (Victor/France/UG).

For community engagement: "I think it's an important way to socialize with the people here or some exchange students and you know, I know some many Finnish bloggers so I know more about some jokes or stories about Finland and I can attend some activities if I wish to, I can see if there are some activities in Tampere or somewhere I can attend to. So yeah, I think it's necessary for me to browse Facebook." (Sally/China/UG).

To meet new people: "One way that I meet new people is to go to a lot of events, for example, Tribe Tampere, they organise many activities and I go there and have random chats with people and then I meet them again randomly on the street or on the bus, gradually getting to know them but getting to know them doesn't mean that we can be friends. But Facebook is a very important channel that brings you information about these events." (Thi/Vietnam/PG).

As a source of information: "If I want to find a group or a service, [Facebook] is a place that I first come to search for. Then… All sorts of other things. Second hand stuff and people moving services, and that sort of things. So it's been really helpful" (Catherine/Vietnam/PG).

To engage with host cultures: "I can join the events which are Finnish. Like I just went to Fazer chocolate factory, it's an event on the Facebook, and while I go there, I just know the thing, the some more details about Fazer chocolate factory and I can just join these events like jogging in the forest and yeah, it make me more closer to the Finnish culture I think." (Camilla/China/UG).

Their various use of Facebook shows the importance of social media as a support tool for migrants. The limited resources offered to them directs these international students to seek out their own sources of information and support; Facebook being well-positioned to fulfil this role. Facebook offers them spaces of their own, which can be utilised for a multitude of functions, dependent on their respective needs.

However, what was also evident in the data collected was the importance of Facebook as a space to navigate cultural identities. Analysis of the profile pages constructed by international students on Facebook demonstrated how they drew on elements of different cultures to curate hybridised identities within their profile pages. The different functions available to users on its platform are set up to enable international students to construct and curate identities for themselves within this third space, addressing their diverse and multicultural audience. These functions become welded as tools through which users were able to create and express identities for their selves, drawing from elements of both their home and host cultures to address their multicultural audiences, creating their own meanings and identities in the process. Their use of Facebook reflects the hybridised positions that international students occupy in their host societies and offers hybridity as an alternative to study migrant experiences.

5 ACKNOWLEDGMENTS

This research was conducted while under scholarship from the University of Western Australia and while on the Finnish Government Scholarship. Many thanks goes out to Rob Cover and Loretta Baldassar for their supervision, help and support.

REFERENCES

Ardévol E. (2012). "Virtual/Visual Ethnography: methodological crossroads at the intersection of visual and Internet research", in Pink, S. (ed.) Advances in Visual Methodology. Sage, London.

Baldassar L., and McKenzie L. (2016). "Beyond "Just Being There" Teaching Internationalization at Home in Two Qualitative Methods Units". Teaching Sociology, 44(2): 84-95.

Baldassar L., and Merla L. (2014). Locating transnational care circulation in migration and family studies. Routledge, London.

Beck U., and Beck-Gernsheim, E. (2002). "A life of one's own in a runaway world: Individualization, globalization and politics", in Beck, U., & Beck-Gernsheim, E. (eds.) Individualization. London: Sage.

Berry J. W. (1997). "Immigration, acculturation, and adaptation". Applied Psychology: An International Review, 46(1): 5-68.

Bhabha H. K. (1994). The location of culture. Psychology Press.

Chen L. and Hanasono L. K. (2016). "The effect of acculturation on Chinese international students' usage of Facebook and Renren". China Media Research, 12(1): 46-59.

Del Pilar J. A. and Udasco J. O. (2004). "Deculturation: its lack of validity". Cultural Diversity and Ethnic Minority Psychology, 10(2): 169.

Fechter A. M. (2007). "Living in a Bubble', in Amit, V. (ed.) Going First class?: New approaches to privileged travel and movement. Berghahn Books, New York.

Gomes, C. (2015). "Negotiating everyday life in Australia: unpacking the parallel society inhabited by Asian international students through their social networks and entertainment media use". Journal of Youth Studies, 18(4): 515-536.

Gomes C., Berry M., Alzougool B. and Chang S. (2014). "Home away from home: International students and their identity-based social networks in Australia". Journal of International Students, 4(1): 2-15.

Kashima E. S. and Loh E. (2006). "International students' acculturation: Effects of international, conational, and local ties and need for closure". International journal of intercultural relations, 30(4): 471-485.

Kim K., Yun H. and Yoon Y. (2009). "The Internet as a facilitator of cultural hybridization and interpersonal relationship management for Asian international students in South Korea". Asian Journal of Communication, 19(2): 152-169.

Montgomery C. and McDowell L. (2009). "Social networks and the international student experience: An international community of practice?". Journal of Studies in International Education, 13(4): 455-466.

Ngo V. H. (2008). "A critical examination of acculturation theories". Critical Social Work, 9(1): 1-6.

Phinney J. (1996). "When we talk about American ethnic groups, what do we mean? American Psychologist", 51(9): 918-927.

Pink S. (2013). "Doing visual ethnography". Sage.

Rudmin F. W. (2003). "Critical history of the acculturation psychology of assimilation, separation, integration, and marginalization". Review of general psychology, 7(1): 3-37.

Sawir E., Marginson S., Deumert A., Nyland C. and Ramia G. (2008). "Loneliness and international students: An Australian study". Journal of studies in international education, 12(2): 148-180.

Weinreich P. (2009). "'Enculturation', not 'acculturation': Conceptualising and assessing identity processes in migrant communities". International Journal of Intercultural Relations, 33(2): 124-139.

Data-Driven Edu Chatbots

Donya Rooein
donya.rooein@polimi.it
Politecnico di Milano
Milan, Italy

ABSTRACT

There are many online courses and contents on the web, so each learner can find the best one and choose it. However, sometimes many online courses are poorly accessible due to the limits of the search engines on the web. The advent of intelligent systems, and online Chatbots, in particular, has brought improvement in various fields. Education Chatbots improve communication, increase productivity, and simplify learning interaction. This study aims to provide an intelligent Edu Chatbot with a high level of customization to learners who have different needs. This way, they can find their personalized learning path dynamically and their customized content without too much time and effort. This is precisely what e-learning needs because of the enormous amount of material on the web.

CCS CONCEPTS

• **Information systems → Specialized information retrieval**;
• **Applied computing → E-learning**.

KEYWORDS

Chatbots, E-learning, Artificial Intelligence, Interactive Tutoring Systems, Adaptive Learning

ACM Reference Format:
Donya Rooein. 2019. Data-Driven Edu Chatbots. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3314191

1 INTRODUCTION

In recent years, digital technology has been evolving in various applications, and the interest in research on intelligent systems has become prominent. Design and implementation of Chatbots are one of most popular kinds of digital technologies. A Chatbot is a computer software that is able to answer some users' questions in a session [5]. The word "intelligent" before Chatbot in an interaction specifies the need for intelligent replies rather than just random ones [6]. Chatbots can be used in various applications such as education, commerce, medicine, and entertainment. Among the big IT companies that have developed Chatbots for industry and research can be named Apple Siri, Microsoft Cortana, Facebook M, and IBM Watson [5]

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314191

Building Chatbots in education (Edu Chatbots) that can interact with learners through natural language like in other applications, had great importance and attracted the attention of researchers in recent years. The purpose of this research is to design and implement an educational framework to provide different educational pathways for learners independently from the content of the subject. The goal of this chatbot is to customize the learning pathway for the learners according to the many attributes . These attributes are driven by profiling the learners and their desires of learning.

In the traditional way of learning, instructors teach content despite the learners' background and goals for attending the course. In Online classes, while students can choose their proper course, they still can not go through a course in different learning pathways in different situations. There is room for variety concerning how an online course should offer diverse learning pathways.

In this study, we hope to find a learning framework for learners through Chatbots to enhance personalization in learning. The main focus of this research is modeling learner interaction with Chatbot and content modeling. The goal of this study is to drive the best learning pathway among various learning choices and bring a high level of flexibility and customization to e-learning for users, according to their goals and desires. There are some studies on the modular architecture for Chatbots which allow them to stay the same while swapping the teaching content [1]. Through this approach, we deploy the Chatbot in several ways at a lower cost. Although, this approach has never been developed on real-world problems. In this PhD thesis, we are going to look for solutions in this regard.

2 STATE OF THE ART

There are many studies related to e-learning system that shed some light on the traditional education system drawbacks [2]. Intelligent tutoring systems [1] are computer-based tutoring systems that tailored to the needs of individuals learners [4]. There is a study on possible effects of a conversational agent in an online portfolio in communication and interaction with a user. The collected data during 4 weeks demonstrated the lack of user satisfaction in completing class activities. However, they had positive feedback on virtual assistant as a social companion. The research introduced three main issues for further research: 1) Learner-developed conversational agents, 2) Improve the agents "intelligence" regarding the domain of knowledge, and 3) Further developed virtual characters that can meet users' humanistic and utilitarian expectations [7]. Rossi et al. in 2011 used a Chatbot in an open source learning management system. The system has been used to support online teachers work by developing a specific knowledge base to pedagogical aspects. The mentioned Chatbot could answer more difficult questions better [10].

[1] ITSs

Wong J.S. et al. used cognitive learning terms in online course materials to study the learning process [11]. In 2018, Clarizia F. et al. presented an educational Chatbot modeling in order to manage best answers to students [5]. They designed this model by using natural language processing techniques to give answers to students.

Nenkov et al. introduced a learning management system to model an intermediary between a student and a teacher [9]. In this study, an intelligent Chatbot was used to automate the interaction between student and teacher implemented using IBM Bluemix platform with IBM Watson technology. IBM Watson is a cognitive system that interacts between users and computers that improve human expertise services by using intelligent techniques such as natural language processing, Analytic, and Machine learning in the form of Chatbots [8].

In a recent study on highly personalized Chatbots in education Casola et al. designed an innovative methodology to provide online courses which are close to the learners‘ desires [3]. They called this methodology iMOOC that has helped the learner to find her learning pathway through the material of the online course by considering her background and profiling attributes. The iMOOC is a content structuring for adaptive education. In this study learners‘ features such as background, effort, needs, depth, and type of interest are taken into account to get the most suitable individualized learning experience. In addition, this structure arranged and organized each piece of contents most properly. To address these points and make adaptive education feasible in iMOOC, they determined items in content and tag them to reflect their properties and the subject of content [1]. Some of the key steps to make education content well adaptive are:

(a) itemized items in small chunks
(b) item meta tagging
(c) content tagging

This methodology by combining these steps can offer learner‘s own learning experience.

Politecnico di Milano [2] and IBM [3] as a joint effort have developed innovative technology in data driven Chatbots research area that named iCHAT [1]. In order to demonstrate the interaction between the user and Chatbot, iCHAT has four components. Conversation routes dynamically according to the information flow that shown in the Figure 1. First, each verified user can launch a stream of conversation with Chatbot through iCHAT interface. It is possible for a user to start several streams of conversation during one session. At the end of the session, all streams can be saved with iCHAT Long Term Memory to be resumed in the next session.

3 PROPOSED APPROACH

Due to the popularity of Chatbots in recent years, many studies have investigated the various aspects of this technology. With respect to current studies on Edu Chatbot, this research aims to study potential ways to have a more intelligent Chatbot in educational purpose. A number of motivations exist for relying on this research to design and develop an intelligent Chatbot through better users‘ understanding. According to previous studies on Online Open Courses

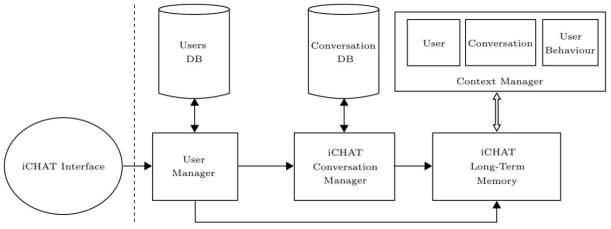


Figure 1: iCHAT‘s overall architecture [1]

(MOOCs), poor rate of completion of an online course is a fundamental challenge [3]. The idea of a model of learners‘ behaviors in the learning process that can identify their learning needs to simplify learning path is to investigate in this study.

This study has concentrated more on the question of understanding the proper learning pathways for each learner, whereas intelligent Chatbot must both understand and generate learning pathways in different conversational streams. Concentrated Study to offer a framework in this regard and implement as an intelligent Chatbot to provide Personalizing to learners is one of the research goals here. Analyzing users‘ data in the learning process is another open challenge in this study to provide a learning pathway dynamically based on the users‘ feedback. There is currently no possibility that you can change your learning pathway of an online course during the course. This feature brings high customization during an online course to the learner and provides her for adjusting the amount of content based on the situation.

Since the goal of learners might be affected by divergent constraints, our most general motivations arise from the fact that it is essential to provide a flexible learning pathway in accordance with these constraints to reduce the probability of not completing the course due to the lack of coordination of the contents of the courses with the student’s needs. There are some features which can enhance Chatbot‘s capabilities to communicate more effectively with users. For instance, we can mention them as an animated gesture supporting knowledge base, provide content from other knowledge sources, and multiple conversational strategies. Besides, by using technologies such as linguistic tools Chatbot enhance the language capabilities to be more like humans.

Another key aspect of this research will be the automation of the possible process in this methodology that is not met in the current process. The quality of the courses and overloads to tag them should be automated as much as possible to save time and accuracy of features related to them. So, this is necessary to model contents properly in this issue. In this respect, content modeling plays an important role in making the Chatbot intelligent. The framework in which the content is correctly modeled increases the accuracy of the Chatbot‘s responses to the user. The general approach for modeling this study is in Figure 2. The novelty concerning current approaches lies in the facet that Chatbot rather than being an expert on the subject is an expert on a body of content covering a subject. In other words, the Chatbot is tasked with efficiently and emphatically interacting with the learner.

[2]HOC-LAB
[3]The Italian research division

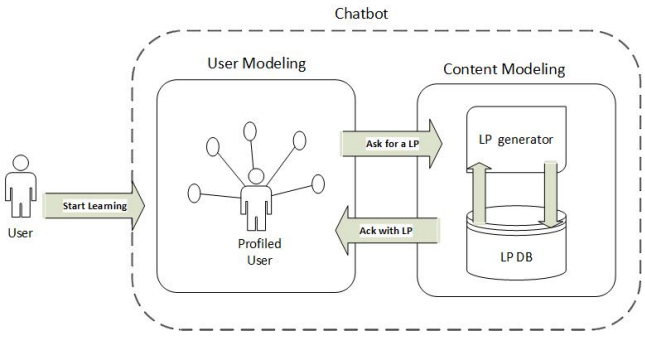


Figure 2: General overview of purposed framework for Chatbot

4 METHODOLOGY

We are currently developing a conversational framework that integrates the real-time aspect of the traditional education system between students and teachers and drive a customized learning pathway from user's profiling. We believe that such a system provides a robust platform on which to continue the development of online education with Chatbots in both academic and industry contents.

The primary research method for this study is to investigate real-world problems in both academic and industry and survey on literature review and conceptual modeling for Chatbots, with the specific focus on Edu Chatbots. A good depth research background is essential to find challenges and research gaps in the problem statement.

In the second stage of this study, there are two sides of the problem statement which should investigate. On the learner side, user profiling identification and classification through a structured approach is at a very beginning stage toward a user modeling. This study will review various types of user intent modeling and their characteristic. In this stage, the iChat architecture will evolve to consider as a part of a framework. On the other side, a study on content modeling is essential to find an efficient structured data on the content of the online courses. A classification method will develop to categorize constraint factors. For the purpose of design learning pathways and modeling, a semantic extraction approach in the industry is essential. Design framework will identify based on a comprehensive review of current industry practices and academic researches.

Third, we should validate our design through design validation approaches such as prototype and get assistance from expert opinions and apply our design in a real context to see the results. There is currently a prototype of iCHAT technology and this work will developing iCHAT to improve quality and adaptive learning in the context of both academic and industry.

Forth, develop and implement the Chatbot into the real online courses. Once our Chatbot is ready, the analytics begin with the collection of data. The data collection is a critical step in providing information due to answer the research questions.

Finally, all the effort and resource dedicated to the previous stages of methodology culminate in the ultimate step. With data, we can analyze whether this approach answers the research questions or not. Evaluate the implementation is in this step. It is possible to measure the rate of the online course completion and compare it with traditional e-learning techniques to evaluate this Edu Chatbot. A quantified approach to measure user engagement is the average time that she spent in a defined period on Chatbot.

During the first year of Ph.D. candidature, the project plan is:

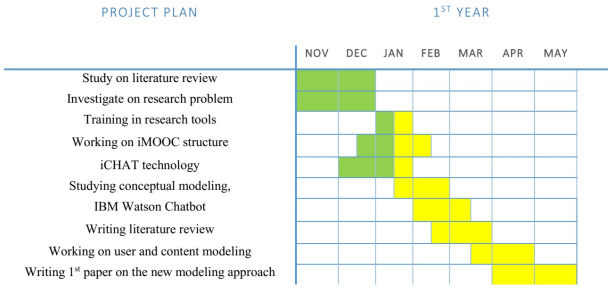


Figure 3: Project Plan

5 RESULTS

This research aims to make possibility into reality in Edu Chatbots context. Edu Chatbot can be used as a tutor, a student evaluator, for a question and answering or to communicate with a teacher. The possible result of this research might be to implement the Chatbot what can easy adapt to the current subject or course so that content manager could do without the requirement of development skills. Moreover, the ultimate Chatbot might be able to bring customization in a personal way by profiling learners. The Chatbot capabilities can expand by including it in other systems such as e-learning systems, a virtual environment or library system or by adding augmenting technology like text-to-speech technology, linguistic tools or animation.

Up to now, after passing two months of my doctoral thesis and studying on this topic, the necessary information has been collected from previous studies. By using this information and the insight of current technologies, the main question has identified for this research and how the outcome of this research can be practical and contribute Chatbot in the context of education. After being able to understand the user status as a set of features. The next introduced feature is the topology navigation. It is essential to understand the user's will, on the other hand, is important to provide a learning pathway that can adapt to those needs. To evaluate one learning pathway which has assigned to a learner, we can consider user's preference of content type, as well as feedback of what a learner did in the previous session such as the difficulty of the last exercise. Then use that information as a prior to create next learning pathway.

6 CONCLUSIONS AND FUTURE WORK

The study aims to further develope the educational Chatbots by reviewing previous works and working on current iCHAT technology and implementing it in the real world application. This research

will be done by finding out the place of Chatbots in education and complete Chatbot requirements in the education context to add pedagogical value in this research area.

A separate line of future research might be to look more at what the Chatbot could bring to the virtual world and if this could become a valuable pedagogical tool. Chatbot with the ability to individualize a service might go beyond in just education context. This idea examines the generalization of this Chatbot to the diverse application such as entertainment and treatment.

As one of the most important motivations for implementing Data Driven Edu Chatbot comes from using current iCHAT technology and iMOOC structure to design and implement a general framework in education and evaluate the efficiency of those purposed approaches.

In the end, we can categorize the current challenges to implement this Edu Chatbot in three major categories as follows:

- Drive important intents to implement in Chatbot as many possible capabilities which can support them (User modeling).
- Finer control on the processing of a sentence. It is essential to measure adjectives in a sentence for adaptive learning by mapping sentences into the specific entities (Content modeling).
- It should be considered in Chatbot to adjust with different styles. Talking in many different styles for Chatbots can be a plentiful challenge if the user interacts with the Chatbot with a small sentence. So, it is better to expand the Chatbot dialog tree in a more human way (Adaptive approach).
- Despite limitations in interacting with the interface, tools can be built to bridge the production gap and put our Chatbot into the efficient way which can reduce time and cost (efficiency).

7 ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my advisors Professor Dr. Barbara Pernici and Dr. Paolo Paolini, you have been a tremendous mentor for me. I would like to thank you for encouraging me to start my Ph.D. in this research field and for allowing me to grow as a research scientist.

REFERENCES

[1] Damla Ezgi Akcora, Andrea Belli, Marina Berardi, Stella Casola, Nicoletta Di Blas, Stefano Falletta, Alessandro Faraotti, Luca Lodi, Daniela Nossa Diaz, Paolo Paolini, Fabrizio Renzi, and Filippo Vannella. 2018. Conversational Support for Education. In *Artificial Intelligence in Education*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 14–19.

[2] Xabier Basogain, Miguel Ángel Olabe, and Juan Carlos Olabe. 2017. Transition to a Modern Education System through e-Learning. In *Proceedings of the 2017 International Conference on Education and E-Learning*. ACM, 41–46.

[3] Stella Casola, Nicoletta Di Blas, Paolo Paolini, and Giuseppe Pelagatti. 2018. Designing and delivering MOOCs that fit all sizes. In *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 110–117.

[4] Sanket Dinesh Choksey. 2004. Developing an affordable authoring tool for intelligent tutoring systems. (2004).

[5] Fabio Clarizia, Francesco Colace, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2018. Chatbot: An Education Support System for Student. In *International Symposium on Cyberspace Safety and Security*. Springer, 291–302.

[6] O. V. Deryugina. 2010. Chatterbots. *Scientific and Technical Information Processing* 37, 2 (01 Apr 2010), 143–147. https://doi.org/10.3103/S0147688210020097

[7] Aaron Doering, George Veletsianos, and Theano Yerasimou. 2008. Conversational agents and their longitudinal affordances on communication and interaction. *Journal of Interactive Learning Research* 19, 2 (2008), 251–270.

[8] Judith S Hurwitz, Marcia Kaufman, and Adrian Bowles. 2015. *Cognitive computing and big data analytics*. John Wiley & Sons.

[9] Nayden Nenkov, George Dimitrov, Yuriy Dyachenko, and Katerina Koeva. 2016. Artificial intelligence technologies for personnel learning management systems. In *Intelligent Systems (IS), 2016 IEEE 8th International Conference on*. IEEE, 189–195.

[10] Pier Giuseppe Rossi and Simone Carletti. 2011. MAPIT: a pedagogical-relational ITS. *Procedia Computer Science* 3 (2011), 820–826.

[11] Jian-Syuan Wong, Bart Pursel, Anna Divinsky, and Bernard J Jansen. 2016. An analysis of cognitive learning context in MOOC forum messages. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1315–1321.

Towards Better Entity Linking Evaluation

Henry Rosales-Méndez
DCC, University of Chile
Supervised by: Aidan Hogan and Barbara Poblete

ABSTRACT

The Entity Linking (EL) task is concerned with linking entity mentions in a text collection with their corresponding knowledge-base entries. Despite the progress made in the evaluation of EL systems, there is still much work to be done, where this Ph.D. research tackles issues concerning EL evaluation. Among these issues, we stress (a) the lack of consensus about the definition of "entity" and the lack of evaluation metrics that allow for different notions of entities, (b) the lack of datasets that allow for cross-language comparison, and (c) the focus on evaluating high-level systems rather than low-level techniques. By addressing these challenges and better understanding the performance of EL systems, our hypothesis is that we can create a more general, more configurable EL framework that can be better adapted to the needs of a particular application. In the early stages of this PhD work, we have identified these problems and begun to address (a) and (b), publishing initial results that constitute a significant step forward in our investigation. However, there are still further challenges that must be addressed before we reach our goal. Our next steps thus involve proposing a more fluid definition of "entity" adaptable to different applications, the definition of quality measures that allow for comparing EL approaches targeting different types of entities, as well as the creation of a customizable EL framework that allows for composing and evaluating individual techniques as appropriate to a particular task.

CCS CONCEPTS

• **Information systems** → *Information extraction.*

KEYWORDS

Information Extraction; Entity Linking; Quality measurement

ACM Reference Format:
Henry Rosales-Méndez. 2019. Towards Better Entity Linking Evaluation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3314199

1 PROBLEM

Entity Linking (EL) is a task in Information Extraction (IE) that focuses on linking the entity mentions in a text collection with entity identifiers in a given Knowledge Base (KB). Such a task has various applications, including semantic search, document classification, semantic annotation, and text enrichment, as well as forming the basis for further IE processes. Despite the fact that works addressing the EL tasks have been pursued by various communities and published in various international conferences, some fundamental questions remain open regarding the aim of the task and how EL results should be evaluated.

A [second]^{t}, larger and more [theatrical]^{t} [Cirque]^{t} [show]^{t}, [{Michael Jackson}^{btdf}: One]^{bt}, designed for [residency]^{t} at the [Mandalay Bay]^{btdf} [resort]^{d} in [Las {Vegas}^{b}]^{td}.

Figure 1: Output annotations of Babelfy (b), TagME (t), DBpedia Spotlight (d) and FRED (f) over the same input

First and foremost, despite the presence of various gold standard datasets, evaluation frameworks, etc., it is still unclear what EL systems should link. There is evidence of disagreement in the EL community on this matter, and as a consequence, different systems target different types of entities. This phenomenon is illustrated in Figure 1, which contains the results for a short example text of four state-of-the-art systems that are popular in the community: Babelfy [23], DBpedia Spotlight [18], TagMe [7] and FRED [9]. As we can observe in Figure 1, there are signs of fundamental disagreements among the involved systems. While FRED and Babelfy consider only proper names, TagME and DBpedia Spotlight also include other nouns for which a corresponding KB entity exists and which do not constitute names. Furthermore, overlapping mentions (denoted by "{}") are targeted by Babelfy and TagME, but not by DBpedia Spotlight nor FRED. So which system is more correct?

This lack of consensus affects further processing of EL systems' outputs since different application scenarios have different requirements on what mentions should be involved. Furthermore, this problem also complicates EL assessment because we do not know how we can define the ideal result that such a system should achieve. Some efforts have been made to standardize which mentions we should identify for annotation, as is the case of the work by Jha et al. [15], who propose a set of rules to serve as guidelines for benchmark creation. However, these rules force the adoption of some considerations that may not suit certain applications and on which there is thus no consensus. For instance, Jha et al., advocate for the omission of overlapping mentions like "{Michael Jackson}", but authors such as Ling et al. [17] disagree. In a semantic search scenario, for example, looking at Figure 1, should such a document be considered relevant for a user interested in texts about Michael Jackson, or more generally, texts about American pop singers?

Several EL benchmark datasets have been proposed that – although used by a variety of systems – also exhibit this disagreement. While KORE50 [12] only annotates proper names, the DBpedia Spotlight dataset [18] includes annotations of common nouns

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314199

such as *software* and *owner*. Additionally, the DBpedia Spotlight dataset includes overlaps, for example, "Google car" is linked with `dbr:Google_self-driving_car`[1], while 'car' is linked to `dbr:Car`.

The solution thus far to address the lack of a consensus has been to define a new consensus, but we propose that a different approach is needed: that no one size fits all in terms of EL. This should not only be reflected in EL systems, but also in EL datasets and metrics used to evaluate EL systems with respect to such datasets. This presents a major challenge tackled in the context of this PhD work.

Such disagreement on what EL systems should link is not the only issue we have encountered in terms of evaluating the EL task. Another major issue is that despite some recent developments for other languages, most work has focused on English texts, both in terms of EL systems and EL datasets. Focusing on a multilingual context, some authors have proposed approaches with a large list of languages. For instance, this is the case of MAG [24], a multilingual EL system that supports annotations over 40 languages. However, the presence of such systems raises new challenges for evaluating the EL task. Such questions now include: How well do EL systems perform outside of English as a primary language? Do multilingual approaches behave equally for all of their supported languages? If not, why not? Are multilingual EL approaches really necessary with recent improvements in machine translation? These questions are not deeply studied yet in the literature. Indeed, only a few of the current EL datasets are multilingual, which complicate any kind of multilingual experimentation.

Generalizing these issues, different applications for EL may bring with them different requirements, which may be best addressed using different techniques. Aside from the issue of the types of entities and the languages targeted, there are also issues regarding for example the length of the text(s), the noise present, the domain of a text, the need to perform EL on semi-structured inputs (e.g., HTML), etc. Different EL systems proposed in the literature have been proposed to solve individual tasks. But individual systems may package together a specific set of techniques, where evaluation is conducted at the level of systems (or ensembles of systems) without understanding which techniques work best in which scenarios. Our ultimate goal, then, is to develop a EL framework that allows for composing and evaluating individual EL techniques, allowing to find the configuration best suited to a particular setting.

2 BACKGROUND

Entity Linking is a task in Information Extraction that focuses on linking the entity mentions in a text collection with entity identifiers in a given knowledge base. Formally, let E be a set of entities in a KB and M the set of entity mentions in a given text collection. The EL process focuses on linking each entity mention $m \in M$ in a text collection with an entity identifier $e \in E$ in a given Knowledge Base (KB). Generally speaking, EL models are commonly separated into two main phases, detailed below:

Entity Recognition (ER) This phase spots which phrases of the input text should be taken as mentions. This problem is also addressed by the Named Entity Recognition (NER) task, where a variety of techniques have been employed to this goal. On the other hand, some works regard ER itself as an independent task, out of the scope of EL [26].

Entity Disambiguation (ED) This phase decides which KB entities should be associated with the identified mentions. This phase is commonly divided into the following steps:

Candidate entity generation: For each entity mention $m \in M$ this stage selects E_m: a candidate set $E_m \subseteq E$ that represents entities with a high probability of corresponding to m is selected. Often this selection is based on matching m with entity labels for E in the knowledge base.

Candidate entity ranking: Each entity $e_m \in E_m$ is ranked according to an estimated confidence that it is the referent of the textual mention m. This can be performed considering a variety of features, such as the perceived "popularity" of e_m, its relation to candidates for nearby mentions, and so forth. The candidate in E_m with the best ranking may be selected as the link for m, possibly assuming it meets a certain threshold confidence (or other criteria).

Unlinkable mention prediction: Some tools consider unlinkable mentions, where no entity in the knowledge base meets the required confidence for a match to a given entity mention m. Depending on the application scenario, these mentions may be simply ignored, or may be proposed as "emerging entities" – annotated as NIL – that could be added to the knowledge base in the future.

In some more recent EL systems, the division between the EL and ED phases is less clear. Some systems apply an End-to-End approach, while other systems apply ER and EL jointly in the same model in the goal of optimizing for both tasks in one process [25, 37]. Other systems assume that entity mentions have already been identified by an existing ER approach and specifically address ED [26].

Several EL approaches have been proposed in the literature. Some of them take annotations, KB entities and their relationship as a graph and perform heuristic to find the proper matching. For instance, Babelfy [23] and AIDA [14] search the densest sub-graph applying a Random Walk with Restart and a greedy algorithm respectively. Other approaches as TagME [7] , THD [4], DBpedia Spotlight [18] and FREME [33] are based on similarity functions between mentions and the KB content. For instance, TagME ranks the candidate entities by two functions: *commonness* and *relatedness*, the first count how frequently an anchor text is linked to a particular Wikipedia entity and the second, returns how often candidate entities for different mentions are annotated from the same Wikipedia page. On the other hand, WIKIME [34] uses multilingual embedding which is trained for words and Wikipedia titles.

3 PROPOSED APPROACH/RESULTS

The present Ph.D. work proposes to address a variety of open questions regarding the evaluation of EL systems. In this direction, the following research questions are being or will be addressed:

(1) What should Entity Linking link?
   (a) How can we define the goal of the EL task?
   (b) Is consensus possible on the definition of an "entity"?
   (c) If not, how can we define benchmark EL datasets and what metrics can we use to reflect the lack of consensus?
(2) How well do EL systems perform in multilingual settings?

[1]Throughout, we use well-known prefixes according to http://prefix.cc

(a) How can we compare EL performance across langauges?
(b) How do EL systems perform for different languages?
(c) Why do results differ across languages?
(d) Are multilingual EL approaches necessary with recent improvement in machine translation?

(3) How can we adapt and configure EL techniques for different applications?
(a) What are the specific applications for EL?
(b) What are the different settings that can be considered?
(c) Which techniques work best under what assumptions?
(d) How can existing EL techniques be best configured to meet the needs of a particular application setting?

We now discuss these three high-level research questions in more detail, describing the issues faced, as well as the ongoing work and plans for future work to address them.

3.1 Lack of consensus

The concept of "named entity" was first coined by the 6th Message Understanding Conference [10] (MUC-6) where entities are assigned to one of the classes *Person*, *Location*, *Organization* and other numeric/temporal expressions. Hence the definition of an entity follows from these classes: any instance of such a class is considered an entity, and per the consensus of MUC-6, the goal of NER is to identify mentions of entities of one of these classes.

With the advent of large-scale, diverse KBs, interest grew in the EL task, which not only identifies (and types) entities in a text, but also links them to the KB. Many ER models and ER benchmark datasets built upon the extensive work in the NER community, proposing to recognize the same entities from the same classes but additionally link them to the KB. This perspective is inherited by many EL systems which continue to identify mentions with NER tools. However, the KBs to which EL systems link often contain classes not considered in the MUC-6 consensus.

Hence EL systems began to develop custom ER techniques that target a broader range of entity types. For instance, in Figure 1, "Michael Jackson" would belong to the MUC-6 class *Person*, whereas "Michael Jackson: One" – though present as an entity in KBs such as DBpedia and Wikipedia – would be excluded by MUC-6, referring to a theatrical production. Along these lines, some authors chose to extend the initial MUC-6 classes, including also *Products*, *Financial Entities* [21], *Films*, *Scientists* [6], etc. On the other hand, other authors propose to separate current classes to more specific ones, for instance, deriving *City*, *State*, *Country* from the class *Location* [8]. Generally speaking, however, class-based definitions of entities are inflexible as they cannot hope to adapt to the variety of types present in large KBs. For instance, Wikidata alone has entities from 50,000 unique classes. Therefore, other authors advocate for more general definitions, but these often lack formality [5, 35]. One option is to use a *Miscellaneous* class of entities, but this leaves the question of what sorts of entities this class should cover. Other authors have tried to provide a more general definition of entity, such as the definition "*substrings corresponding to world entities*" used by Ling et al. [17]; however, such a definition is cyclical, due to using the word "*entity*" in the definition of an "*entity*".

Table 1: Survey of popular EL datasets; for multilingual datasets, the quantities shown refer to the English data available. We present metadata about the relaxed and strict version of our dataset by VoxEL$_R$ and VoxEL$_S$ respectively.

Dataset	Languages
AIDA/CoNLL-Complete [11]	EN
KORE50 [13]	EN
IITB [16]	EN
ACE2004 [27]	EN
AQUAINT [27]	EN
MSNBC [3]	EN
DBpedia Spotlight [19]	EN
N3-RSS 500 [28]	EN
Reuters 128 [28]	EN
Wes2015 [36]	EN
News-100 [28]	DE
Thibaudet [1]	FR
Bergson [1]	FR
SemEval 2015 Task 13 [22]	EN,ES,IT
DBpedia Abstracts [2]	DE,EN,ES,FR,IT,JA,NL
MEANTIME [20]	EN,ES,IT,NL
VoxEL$_R$	DE,EN,ES,FR,IT
VoxEL$_S$	DE,EN,ES,FR,IT

Proposal: Instead of addressing the abstract question "*what is an 'entity'?*", our position is to rather address the more practical question "*what should Entity Linking link?*". Posed this way, the question suggests a practical response: "*it depends on the application!*" [32]. For example, in a semantic search scenario, where the goal is to find documents mentioning particular entities or particular types of entities, finding all repeated mentions of an entity may not be so key a requirement for EL: finding any mention in the document might suffice. On the other hand, for relation extraction, each mention might refer to a potential relation in the text, and hence the requirements for EL change. With this in mind, our goals are to first understand on which types of entities there is consensus in the community, and on which not. Then we wish to better understand what are the applications for EL, and how the choice of application affects the requirements of the EL system. Finally we aim to develop EL benchmarks and metrics that – rather than assuming a one-size-fits-all definition of an entity – reflect our findings in terms of varying consensus, applications and requirements.

3.2 Multilingual EL

Thus far, the bulk of effort in EL research has been devoted to English texts. However, more recently, a number of multilingual EL systems – supporting multiple languages – have been proposed. Such systems raise new questions for evaluating EL: How well would state-of-the-art approaches perform over non-English corpora? How would performance vary across languages (and why)? Given that EL often targets named entities, how important is it for EL systems to be configurable for different languages (e.g., *Michael*

Table 2: Overall EL evaluation (F_1) of selected approaches for the SemEval 2015 Task 13 in Spanish (ES) and English (EN). Approaches configured for Spanish are italicized.

System	ES	EN
Babelfy	0.439	0.602
DBpedia-Spotlight	0.337	0.414
WikiMe	0.043	0.043
TAGME	0.133	0.395
THD	0.069	0.110
AIDA	0.010	0.046

Jackson's name does not change with language, only with alphabet)? Given recent improvements in machine translation, how do multilingual EL systems perform versus translating input text?

One challenge faced for responding to these questions is the short list of multilingual datasets available that could be used to evaluate EL performance for various languages. In Table 1 we survey a variety of EL datasets available in the literature, where we can observe in Table 1 that the majority only consider English text; others that consider non-English texts only offer one language. While a number of multilingual datasets have now been made available, they further present some issues for comparing EL systems across languages, as we will discuss later.

Initial Results: Taking an existing multilingual dataset – SemEval 2015 Task 13 – in [31], we perform initial experiments to compare the performance of popular EL systems for English and Spanish texts, testing Babelfy, DBpedia-Spotlight, WikiMe, TagME, THD and AIDA. Only the first three of these systems can be explicitly configured for Spanish texts. Table 2 offers an overview of the main results. As can be observed, systems generally perform considerably better for English than Spanish, particularly (but not limited to) EL systems not configurable for Spanish. We consider this result as being potentially due to three main factors (a) KBs (e.g., Wikipedia) contains different information for both languages with potentially more information available in English, (b) the models/techniques change according to the target language where, for example, DBpedia-Spotlight's ER uses different models according to the targeted language, and (c) the presence of variations in the languages themselves, where, for example, recognizing "*Star Wars*" is less challenging than the Spanish version "*La guerra de las galaxias*" due to capitalization rules in Spanish and the phrase length.

Further Results: One of the obstacles to ongoing research on multilingual EL is the low availability of datasets with the same text in different languages. According to our survey in Table 2, there are only three multilingual datasets available (the VoxEL datasets are proposed by us). As we detected in the initial work described previously, each of the three datasets has its own limitations. SemEval 2015 Task 13 is composed of four documents on biomedical, math, computer and social topics; DBpedia Abstracts[2] is a large corpus build automatically from the abstracts (first paragraph) the Wikipedia pages, containing in total 39132 documents; and MEANTIME contains annotations of 120 news articles from WikiNews[3] with annotations of entities, events, temporal information and semantic roles. However, DBpedia Abstracts is not a parallel corpus: the text differs across languages, making it unsuitable for comparing performance across languages. On the other hand, while SemEval 2015 Task 13 and MEANTIME aim to be parallel corpora, they have different annotations in different languages. Hence these datasets are not ideal for comparing the performance of EL systems across different languages (though they can be used for comparing EL systems across individual languages). Furthermore, these datasets adopt a particular notion of entity, which as argued previously, may not be that agreed upon by the community.

In order to better compare EL performance across multiple languages, we proposed a new multilingual corpus called VoxEL [30] with the aim of ensuring the same annotations across different languages, as well as reflecting in the dataset the lack of consensus on what is an entity. VoxEL is based on 15 news articles from the European newsletter VoxEurop[4], which is translated by professionals to different European languages. We first aligned the sentences and entities across languages, resolving cases where some entities and sentences were omitted/changed in the translations. To address the lack of consensus about what is an entity, we include two versions of the same dataset: one *strict* version that includes only those entities that all systems appear to agree should be linked (i.e., *Person*, *Location* and *Organization*), and another *relaxed* version that includes also all mentions (including overlapping mentions) with a Wikipedia page related to them (e.g., 'software', 'resort', etc.). Our results again show that – aside from Babelfy – EL systems generally perform much better over English texts. We also compare in [29] the idea of using machine translation in EL environments to translate the input text rather than configuring the EL system for the native text. The results show that the majority of systems – namely DBpedia Spotlight, FREME and TagMe – perform better when the input text is either in English, or translated to English.

3.3 Configurable EL Framework

Our general hypothesis in this Ph.D. work is that when it comes to EL systems, one size does not fit all: different scenarios and different applications may have different requirements for an EL system, including, but not limited to, the types of entities targeted, the languages supported, etc. Drawing the Ph.D. work together, our goal is to be able to perform finer-grained evaluation of EL systems under different requirements and different assumptions. Going one step further, rather than evaluating EL systems, we wish to be able to evaluate the effects of different ER and ED techniques. A given EL system already configures a number of techniques into one solution, where the results of a system then confound the performance of these techniques. While some EL papers present results for ER and ED tasks separately, we propose to go one step further.

Proposal: We propose to create a modular EL framework that implements a selection of the most important EL techniques found in the literature. Such a framework should allow for creating custom EL pipelines that allow for evaluating and composing techniques

[2]http://wiki-link.nlp2rdf.org/abstracts/; January 1st, 2018

[3]https://en.wikinews.org/; January 1st, 2018

[4]http://www.voxeurop.eu/; January 1st, 2018

according to the requirements of a particular application. The challenges posed by this proposal are various and include not only the engineering challenge of developing such a modular framework but also some non-trivial research questions. First we must address the question of how the requirements of a particular application can be represented. Second we must consider what datasets and metrics can be used to evaluate individual EL techniques (and their composition) per these requirements. Though ambitious, if successful, this stage of the Ph.D. work could lead to a better understanding of how EL techniques perform under different assumptions and further offer the practical contribution of a generalized EL framework adaptable to a broader variety of applications and settings.

4 METHODOLOGY

Our next steps are aimed at understanding the consensus in the community regarding what is an "entity" in the context of EL, and defining evaluation protocols (datasets and metrics) accordingly. Further into the future, we aim to start work on the configurable EL framework. More specifically, the steps currently in progress are: (1) Review the state-of-the-art in measures used for EL evaluation. (2) Issue a questionnaire to the EL community to understand the current consensus on the goal of EL systems. (3) Propose a categorization of entities that allows for the inclusion/exclusion of entities per the application requirements. (4) Propose finer-grained evaluation protocols for EL according to previous findings. Later we further wish to being to: (5) Implement a general EL framework based on the previous results that allows us to evaluate and combine techniques according to the application scenario, comparing results against state-of-the-art EL systems in different settings.

5 CONCLUSION

This Ph.D. proposal is motivated by the hypothesis that for EL systems, one size does not fit all. Our first aim has been to understand how the goal of EL systems may vary across different applications and how that affects the consensus of what is an "entity"; this work remains ongoing. Our second aim is to consider how different EL systems perform for different languages, where we have published some results and proposed a novel dataset along these lines. Our third aim is to develop a general EL framework that allows greater configurability for a given application and setting. Our overall ambition is to reach a greater understanding of how EL techniques perform in different settings and arrive at a practical framework that can be used in a broader variety of applications that can, potentially, benefit from the results of the EL task.

Acknowledgments: I wish to thank Aidan Hogan and Barbara Poblete for supervising my PhD research. This work was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017 and by the Millennium Institute for Foundational Research on Data (IMFD).

REFERENCES

[1] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: named entity linking in digital literary editions using linked data sets. *CSIMQ* 7 (2016), 60–80.
[2] Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus.. In *LREC*.
[3] Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *EMNLP-CoNLL* (2007), 708.
[4] Milan Dojchinovski and Tomás Kliegr. 2013. Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In *ECML/PKDD*. 654–658.
[5] Alan Eckhardt, Juraj Hresko, Jan Procházka, and Otakar Smrs. 2014. Entity linking based on the co-occurrence graph and entity probability. In *ERD*. 37–44.
[6] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artif. Intell.* 165, 1 (2005), 91–134.
[7] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*. 1625–1628.
[8] Michael Fleischman. 2001. Automated Subcategorization of Named Entities. In *ACL*. 25–30.
[9] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic Web Machine Reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.
[10] Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *COLING*. 466–471.
[11] Johannes Hoffart and et al. 2011. Robust disambiguation of named entities in text. In *EMNLP*. ACL, 782–792.
[12] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *CIKM*. 545–554.
[13] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *CIKM*. 545–554.
[14] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *EMNLP*. 782–792.
[15] Kunal Jha, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. All that Glitters Is Not Gold - Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking. In *ESWC*. 305–320.
[16] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *SIGKDD*. 457–466.
[17] Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design Challenges for Entity Linking. *TACL* 3 (2015), 315–328.
[18] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*. 1–8.
[19] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*. ACM, 1–8.
[20] A.L. Minard and et al. 2016. MEANTIME, the NewsReader multilingual event and time corpus. (2016).
[21] Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *LREC*.
[22] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking.. In *SemEval*. 288–297.
[23] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL* 2 (2014), 231–244.
[24] Diego Moussallem, Ricardo Usbeck, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2018. Entity Linking in 40 Languages Using MAG. In *ESWC*. 176–181.
[25] Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2016. J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features. *TACL* 4 (2016), 215–229.
[26] Julien Plu, Giuseppe Rizzo, and Raphaël Troncy. 2016. Enhancing Entity Linking by Combining NER Models. In *ESWC*. 17–32.
[27] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *NAACL-HLT*. 1375–1384.
[28] Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In *LREC*. 3529–3533.
[29] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. Machine Translation vs. Multilingual Approaches for Entity Linking. In *ISWC (P&D/Industry/BlueSky)*.
[30] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In *ISWC*. 170–186.
[31] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2017. Multilingual Entity Linking: Comparing English and Spanish. In *LD4IE*. 62–73.
[32] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What Should Entity Linking link?. In *AMW*.
[33] Felix Sasaki, Milan Dojchinovski, and Jan Nehring. 2016. Chainable and Extendable Knowledge Integration Web Services. In *NLP&DBpedia*. 89–101.
[34] Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual Wikification Using Multilingual Embeddings. In *HLT-NAACL*. 589–598.
[35] Victoria S. Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of*

Web Semantics 4, 1 (2006), 14–28.

[36] Jörg Waitelonis, Claudia Exeler, and Harald Sack. 2015. Linked data enabled generalized vector space model to improve document retrieval. In *NLP&DBpedia*.

[37] Longyue Wang, Shuo Li, Derek F. Wong, and Lidia S. Chao. 2012. A Joint Chinese Named Entity Recognition and Disambiguation System. In *CIPS-SIGHAN*. 146–151.

Velocity on the Web - a PhD Symposium

Riccardo Tommasini
supervised by Emanuele Della Valle
Politecnico di Milano, Department of Electronic, Informatic, and Bioengineering
Milan, Italy
riccardo.tommasini@polimi.it

ABSTRACT

It is a streaming world: a new generation of Web Applications is pushing the Web infrastructure to evolve and process data as soon as they arrive. However, the Web of Data is not appealing to the growing number of Web Applications demanding to tame Data Velocity. To solve these issues, we need to introduce new key abstractions, i.e., stream and events. Moreover, we need to investigate how to identify, represent and process streams and events on the Web. In this paper, we discuss why taming Velocity on the Web of Data. We present a Design Science research plan that builds on the state of the art of Stream Reasoning and RDF Stream Processing. Finally, we present our research results, for representing and processing stream and events on the Web.

KEYWORDS

Web of Data, Stream Processing, Complex Event Processing, Stream Reasoning

ACM Reference Format:
Riccardo Tommasini. 2019. Velocity on the Web - a PhD Symposium. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3314192

1 PROBLEM

Data Velocity is the challenge to process data as soon as they arrive, and before it is too late. A new generation of Web applications is raising in a streaming world [41], where there is the need to process streams continuously and detect events reactively. Relevant application scenarios like environmental monitoring [1], social media trend analysis [24], and anomaly detection in financial markets [33] prove that Data Velocity is a critical challenge for the Web [17, 23].

The Web of Data (WoD) is an extension of the World Wide Web (Web) that encourages data sharing and interoperability. The fundamental notion of the Web are Resources to discover, to represent, and to process. WoD builds on the Web infrastructure. It uses URIs to enable discovery by identifying resources; it uses HTTP to enable links following (i.e., de-referencing) and processing.

Velocity is related to the entire Web data infrastructure, e.g., representing data, managing, delivering to recipients, reacting to changes, and responding to failures promptly. The Web underlying infrastructure is evolving accordingly by adding new protocols such as HTTP Long-Polling[1], HTTP Streaming[2], and WebSockets[3] that provide data continuously, i.e., in a *streaming* fashion; Server-sent events[4] and WebHooks[5] that supply data reactively, i.e., in an *event-driven* fashion. Henceforth we call HTTP extended with these new protocols HTTP*.

Web resources are heterogeneous. Therefore, WoD provides principles to enable interoperability: (i) exchange data using the Resource Description Framework (RDF)[6] a machine-readable data model (ii) build vocabularies using proper knowledge representation languages as RDF Schema[7] or the Web Ontology Language[8] (OWL); (iii) share vocabulary to annotate data semantically and (iv) use SPARQL[9] to process data in its broadest sense.[10]

The WoD is an evolving environment [21], but its key abstractions – i.e., RDF Triples, Named Graphs and RDF Datasets – are not adequate to tame data velocity. The WoD community proposed *RDF triples*, i.e., $< subject, predicate, object >$, to publish rich and interconnected Web documents. The community proposed *Named Graph* [10], i.e., set of RDF triples (graphs) identified by URIs, to organize document adding contextual metadata. Indeed, metadata are useful to support trust, syndication, and provenance, and they can boost data processing [18]. The WoD community proposed *RDF Datasets*, i.e., collections of Named RDF Graphs, to support the Linked Open Data (LOD) initiative. Indeed, LOD users need to discover, represent, and process multiple RDF graphs at once [5].

Problem Statement: the Web of Data is currently not appealing to the growing number of Web applications demanding to tame Velocity [7]. To solve this problem, the WoD needs new adequate key abstractions, i.e., *Streams* and *Events* [11]. Moreover, we need to investigate *how to discover, represent, and process* stream and events on the Web.

The research goals implied by this problem requires the design of artifacts that improve the WoD. Therefore, we run our investigation following a design science methodology [45], i.e., the *design cycle*, which is reported in Figure 1 and detailed in Section 3.

2 STATE OF THE ART

Previous attempts to tame velocity on the Web go under the umbrella of Stream Reasoning [17, 23]. Although there are proposals

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314192

[1] https://realtimeapi.io/hub/http-long-polling/
[2] https://realtimeapi.io/hub/http-streaming/
[3] https://www.w3.org/TR/websockets/
[4] https://www.w3.org/TR/2009/WD-eventsource-20090421/
[5] https://www.w3.org/TR/websub/
[6] https://www.w3.org/TR/rdf11-concepts/
[7] https://www.w3.org/TR/rdf-schema/
[8] https://www.w3.org/TR/owl2-overview/
[9] https://www.w3.org/TR/sparql11-overview/
[10] On the Web of Data, processing certainly means querying, but also reasoning, managing data access and storage, orchestration, choreography, and federation.

that discuss how to discover [4, 25, 29] and represent stream and events [26, 30, 44], most of the work focus on processing RDF Streams [9, 15, 16, 23] and Events [3, 13] (RSP). Due to lack of space, we cannot list all the relevant works; we invite the interested readers to consult the recent surveys [14, 23].

Discover. Sequeda and Corcho [29] proposed Linked Streaming Data as a mechanism to identify sensors and their observations. The paper provides a set of requirements and a proposal for URIs referencing space and time. Barbieri and Della Valle [4] propose to identify streams using URIs to enable discovery as for Linked Data. The authors focus on RDF Streams, and they model them using two kinds of named graphs. Each element of the stream is an Instantaneous Graphs (iGraphs), identified by a timestamp. The Stream Graphs (sGraphs) describes the current content of the window over the RDF Stream, and it references the iGraphs using the properties *rdfs:seeAlso* and *receivedAt*.

Represent. The Semantic Web proposals for representing events include, but are not limited to, the Event Ontology [25], Linking Open Descriptions of Events (LODE) [30] and Simple Event Model [44]. However, neither of these focuses on the events as a way to tame velocity. Only Rinne et al. [26] proposal for an ontology design pattern for events has the primary intent of reactive detection. Proposals for stream focus more on the content[11] than on stream as resources. In [4] Barbieri e Della Valle sketches a vocabulary to use for sGraphs. Moreover, Keskisarkka [22] proposed a SPIN extension for RSP which introduces some terminology based on RSP-QL (a reference model for modelling RSP engines). Recently, Schema.org included DataFeed[12] and DataFeedItem[13] as classes to model streams of content.

Process. The generality of the term “processing" calls for more specificity. Our investigation includes: *Querying*, i.e., event detection and continuous transformations as defined in the Information Flow Processing state-of-the-art [11] (IFP)[14]; *Data Management*, i.e., persisting and accessing streams and events, but also enabling orchestration, and choreography of services that produce/consume streams or events; and *Reasoning*, i.e., inference algorithms to enrich streams and derive implicit events.

RSP attempts to tame velocity and variety simultaneously [42] combining solutions from IFP and Semantic Web. RDF Streams are formalized as infinite sequences of pairs (G, t) where G is an RDF Graph and t is a non-decreasing timestamp [16]. Dell'Aglio et al. proposed a reference model called RSP-QL [16] that captures formal and operational semantics of continuous extensions of SPARQL [23]. RSP-QL extends the concepts of RDF Graph and RDF Dataset to enable the continuous semantics. Moreover, it adds stream-oriented operators such as Time Windows.RSEP-QL [13] extends RSP-QL with CEP operators to capture the semantics of RSP dialects for processing Events [2]. It introduces the concept of Basic Event Pattern as an extension of SPARQL basic graph pattern with temporal annotations.

[11] https://www.w3.org/TR/vocab-ssn/
[12] https://schema.org/DataFeed
[13] https://schema.org/DataFeedItem
[14] Due to the lack of space we cannot include a comprehensive background on IFP, the interested reader can consult [11] for a recent survey.

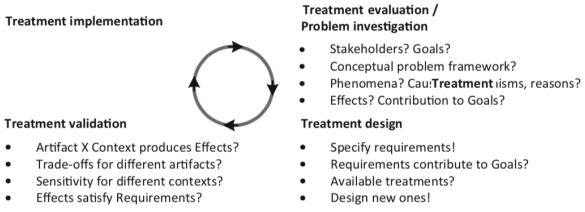


Figure 1: Design Cycle from [45].

Despite the attempts to formalize and uniform the approaches for RSP, existing dialects do not follow a common design methodology. In particular, they do not satisfy language design principles [12].

Stuckenschmidt et al. [32] envisioned the idea of cooperation of stream and event processors. They elaborated on orchestration proposing Cascading Reasoning, i.e., a hierarchical approach that combines multiple stream/event processor into layers of different expressiveness. They elaborated on choreography proposing Networks of Stream Reasoners cooperating on the Web. Finally, Calbimonte et al. [8] further develop this second idea towards a reactive Web Stream Processing. However, none of these visions reached the level of maturity to be validated and evaluated.

Works on reasoning divide into query rewriting for RSP [9] and incremental maintenance of ontology materialization [15]. Both the approaches aim at boosting reasoning performance to meet velocity requirements. Works on query rewriting attempt to extend the mapping languages with the concept of Window. Works on incremental reasoning focus on extending reasoners' capability with IFP concepts, while the idea of introducing reasoning features in an IFP engine is stil an unexplored path.

3 METHODOLOGY

For our investigation, we follow a Design Science (DS) [45] research methodology. DS is the investigation of the interaction of *artifacts*, e.g., algorithms, methods, techniques, with a *problem context* in order to improve it. Such interaction, called *treatment*, is designed by researchers to solve a problem. The application of treatments to the original problem context is called *implementation*.

Artifacts and treatments are designed following a subset of the engineering cycle [45], i.e., the *design cycle*, which is reported in Figure 1. It starts with a research goal. DS research goals divide into *treatment evaluation* and *problem investigation*. The the design cycles continuous with *treatment design* and *validation*.

Problem Investigation happens before implementation and requirements analysis. DS identifies *Design Problems*, which are the problems to (re-)design an artifact so that it better contributes to achieving some goal, and *Knowledge Questions*, i.e., explanatory or descriptive questions about the world.

Knowledge questions do not call for a change in the world, and the answer is a unique falsifiable proposition. On the contrary, design problems call for a change in the world. A solution is a design, i.e., a decision about what to do that is generally not unique.

Treatment Design. Design problems assume a problem context and some stakeholders to interview for deriving *requirements*, i.e., the desired property of the artifact. The researcher has to formulate

contribution arguments, i.e., sustain that an artifact which satisfies the requirements contribute to stakeholder goals.

Starting from a design problem we can ask knowledge questions about the artifact, the context, and the treatment, e.g., we can ask about the performance of the of the artifact. Similarly, knowledge questions can lead to a design problem, e.g., building a prototype of the artifact or simulating the context. These are all *empirical knowledge questions*, which require data about the world to answer them. To this extent, DS identifies *instrument design problems* that are the problem to design an instrument that support the investigation, e.g., a simulation or a prototype.

Treatment Validation aims at predicting how an artifact will interact with its context, without actually observing an implemented artifact in a real-world context. Validation is done under controlled conditions, before implementation, and using a *validation model*, i.e., exposing an artifact model to problem context models.

Validation include asking empirical knowledge questions about i) the effects produced by the treatment; ii) whether or not these effects satisfy the requirements; iii) the trade-offs between different artifacts applied to same problem context, and iv) the sensitivity between the same artifact applied to different contexts.

Research methods used for validation are (i) Expert opinion, which usually corresponds to peer-reviewing. (ii) Single-Case Mechanism Experiments, which studies the response of a validation model to a stimulus under controlled conditions and regarding the model internal. (iii) Statistical Difference-Making Experiments, in which a researcher compares the outcomes of different treatments applied to the same model context, and; (iv) Technical Action Research (TAR), which studies the properties of an artifact under development used by a researcher under real-world conditions.

Treatment Evaluation investigates the treatment as applied by stakeholders in the real world after implementation. Research methods used for evaluation since include (i) Surveys; (ii) Observational Case Studies; (iii) Single-Case Mechanism Experiments and (iv) Statistical Difference-Making Experiments.

4 PROPOSED APPROACH

To extend the Web of Data to tame velocity, we need to prescribe to discover, represent and process Stream and Events as resources.

On the one hand, considering recent extensions to the Web architecture, a discovery approach based on HTTP* and URIs is reasonable [29, 42]. Moreover, we should take into account what done by the Stream Reasoning community [14] on extending RDF and SPARQL query language for processing streams and events.

On the other hand, representing and processing in its broadest sense 10 are still open challenges towards taming velocity on the WoD. Thus, we formulate the following research question.

Research Question: *Can we enable representation and processing of stream and events to tame velocity in the Web of Data?*

Two problems arise from this research question: the representation problem and the processing problem for streams and events.

Moreover, the lack of a systematic approach of empirical research in the field [27] highlights the need for designing a research instrument that can help to answer many empirical knowledge questions. We refer to this as validation problem.

4.1 Representation Problem

Representing streams and events on the Web of Data is a design problem that we formulate as: *improve the Web of Data by designing an artifact, which satisfies Web developers' requirements to represent streams and events.*

WoD prescriptions for knowledge representation are RDFS and OWL2 and, thus, our solution must be compatible with them. Therefore, *we assume that existing knowledge representation languages for the Web of Data can capture streams and events semantics.*

As presented in Section 2, works on stream representation focus on the content, and a shared terminology to describe streams on the Web is still missing. Moreover, several ontological models compete for representing events, but they were designed without taking into account Velocity. Therefore, we plan to

- RP.1 design and share a vocabulary for streams descriptions that satisfies Web developers' requirements.
- RP.2 survey existing event models to propose an artifact that allows event definitions without neglecting Data Velocity [7].

4.2 Processing Problem

Processing streams and events as we discussed in Section 2 is also a design problem that we can formulate as: *improve the Web of Data by designing an artifact, which satisfies Web developers' requirements to process streams and events.*

In Section 2, we highlight how existing solutions for querying streams and detecting events on the Web do not follow common design principles. In particular, stream and events are not first-class citizen in the query languages. Moreover, cooperation of web stream and event processors calls for further investigations. Finally, works on reasoning focused on extending reasoners to meet velocity requirements rather than extend stream and event processing engines with inference capabilities. Therefore, we plan to

- PP.1 re-design an RSP approach that follows common language design principles for querying streams and events [12];
- PP.2 investigate techniques for orchestration and choreography of Web Stream Processors that can cooperatively solve problems in a decentralized Web environment [32].
- PP.3 investigate how to extend event processing languages and event-based architecture with reasoning capabilities to meet the performance requirements for treating velocity [31].

4.3 Validation Problem

The validation problem is an instrument design problem that we can formulate as: *improve empirical research for velocity on the WoD by designing an artifact, which satisfies researcher requirements, for validating artifacts that represent and process streams and events.*

Validation of the representation problem is based on design principles such as Tom Gruber's [20]. However, the for knowledge representation artifacts the improvement to the problem context is quantified by adoption. Technical Action Research (TAR) is a validation method that can (i) foster the artifacts adoption, while (ii) improving the artifact design with a strict feedback loop.

Validation for the processing problems, henceforth referenced as benchmarking, requires instead to measure precise performance requirements [31]. Although there are principles for domain-specific

benchmarking [19], existing benchmarks lack a systematic and comparative research approach. In particular, (i) prototypes might be unavailable and (ii) do not follow common design guidelines; (iii) the manual effort to set up the benchmark is vast, and (iv) the experimental environment does not guarantee reproducibility and repeatability. These issues make it hard to run the necessary validation, i.e., Single-Case Mechanism Experiments or Statistical Difference Making Experiments (see Section 3). Thus, we plan to

- VP.1 design an experimental environment that guarantees reproducibility and repeatability of experimental results;
- VP.2 enable Single-Case Mechanism Experiments by design an artifacts that can reproduce the variety of RSP engines without neglecting common design principles;
- VP.3 enable Statistical Difference Making Experiments by allowing fast-prototyping of alternative validation models and, thus, rapid exploration of the solution space; and
- VP.4 organize TAR sessions to using research instruments to validate representation and processing artifacts.

5 RESULTS

In this section, we discuss how we applied Design Science (DS) research methodology to our research plan. In particular, we focus on how we implemented the design cycle for the representation and processing problems, how we validated and how we evaluated the designed treatments.

5.1 Representation Problem

Regarding the representation problem, we designed a Domain-Specific Language (DSL) for event definition and processing called OBEP [35], and a vocabulary for describing streams called VoCaLS [36]. Following the design cycle, we collected requirements and we validated the treatments by expert opinion respectively for the DSL in [34], and for the vocabulary in [28].

```
<> a vocals:StreamDescriptor ; dcat:dataset :MilanTrafficStream .

:MilanTrafficStream a vocals:RDFStream ;
  vocals:hasEndpoint :MilanTrafficStreamEndpoint ;
  dcat:publisher <www.3cixty.eu>;
  dcat:description "Stream produced by traffic sensors in Milan"^.

:MilanTrafficStreamEndpoint a vocals:StreamEndpoint ;
  dcat:license <https://creativecommons.org/licenses/by-nc/4.0/> ;
  dcat:format frmt:JSON-LD ;
  dcat:accessURL "ws://example.org/traffic/milan".
```

Listing 1: Example of RDF Stream using VoCaLS.

We addressed 4.1 in [36], where we inquired WoD developers about their requirements for stream representation, and we provided evidence of the vocabulary usability. Listing 1 shows an example of stream description using VoCaLS that enables web stream processing scenarios not possible before.

We addressed 4.1 in [35], where we compared our event definition DSL with existing RSP event processing language, i.e., EP-SPARQL and RSEP-QL. We show that our approach, based on Description Logic (DL) reasoning, is more straightforward and more succinct than existing ones. DL is the foundation of OWL2 and, thus, the results respect the assumption made in Section 4. Listing 2 shows an example OBEP event declaration that is based on rules and, thus, can tame Velocity in the form of Event-Condition-Action [7].

```
EVENT :SmokeDetectionEvent subClassOf
   (hasContext some (hasProperty some Smoke))
EVENT :HighTemperaturEvent subClassOf TemperaturEvent and
   (result some (hasValue >= 40))
NAMED EVENT :FireEvent
MATCH :HighTemperaturEvent seq :SmokeDetectionEvent  WITHIN (5m)
```

Listing 2: Example of OBEP Language

5.2 Processing Problem

Regarding the processing problem, we planned to design several artifacts that address different aspects of processing, i.e., querying (4.2), data management (4.2) and reasoning (4.2).

Our investigation related to 4.2 develops in [35, 37, 38]. In [35], we also provide the specification of an event processing algebra that treats events as first-class citizens. Preliminary validation of the approach compares it with existing models to represent events. In [37], we highlighted the issues related to designing a query language based on RSP-QL formalization. A language that treats streams as first-class citizen should keep the languages constructs minimal, homogeneous, symmetric and orthogonal [12]. In [38], we proposed a first artifact prototype for such an RSP-QL engine called YASPER that will support our empirical investigation.

Our investigation related to 4.2 develops in [6], where we showed the feasibility of Cascading Reasoning, i.e., an orchestration approach, envisioned by Stuckenschmidt et al. [32]. The approach we designed combines exiting RSP and Stream Reasoning solutions into a layered architecture. To validate the treatment, we run Single-Case Mechanism Experiments using our prototype.

Last but not least, our investigation related to 4.2 develops in [6, 35] where we investigated how to combine existing stream reasoning approaches using RDF Streams. Moreover, an approach for event-based hierarchical reasoning is under review.

5.3 Validation Problem

Regarding the validation problem, we formalized a systematic and comparative research approach [39] and a Web environment for experimentation [40]. Our approach is based on two notions, i.e., Experiment and Baseline [39]. The former characterizes the model context for the validation, while the latter represents a versatile validation model [45].

To address 4.3, we developed an open-source Web environment called RSPLab [40] that guarantees reproducibility and repeatability of experimental results using vitalization techniques. The researcher needs to know the prototype internals in order to explain its response out of a stimulus to address 4.3. The researcher needs to fast-prototype many alternatives that can be compared systematically [39] to address 4.3.

Finally, we are organizing a series of tutorials to apply TAR with a knowledgeable audience. The initial sections include RW 2018, ISWE 2018, and The Web Conference 2019 [42].

6 CONCLUSION AND FUTURE WORK

In this paper, we discussed why and how extending the the Web of Data for taming Velocity. Our research goal is to allow Web developers to discover, represent and process streams and events.

To achieve it, we presented a design-science research approach that elaborates into a representation problem, a processing problem,

and a validation problem. We designed a vocabulary for stream description called VoCaLS [36] and a DSL for event definition called OBEP [35] to solve the representation problem. We developed our investigation for processing problem in three different directions, i.e., querying [35, 37, 38], reasoning [6, 35] and orchestration [6]. We designed an research instrument [40]and a methodology [39] to solve the validation problem.

Our research plan now focuses on validation and evaluation [45]. We aim at using Technical Action Research to validate VoCaLS and OBEP. To this extent, we are integrating VoCaLS and OBEP within our tutorials [43], and extending RSPLab to track usability metrics.

We plan to design a unified approach for querying that keeps the stream/event operations orthogonal of the remaining part of the algebra. We plan to evaluate such an approach running Statistical Difference-Making Experiments.

An approach that exploits VoCaLS descriptions to make a choreography of Web Stream Processors on a web environment is currently under review. We plan to validate the approach running Single-Case Mechanism Experiments.

REFERENCES

[1] Damminda Alahakoon and Xinghuo Yu. 2016. Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey. *IEEE Trans. Industrial Informatics* 12, 1 (2016), 425–436.

[2] Darko Anicic. 2012. *Event Processing and Stream Reasoning with ETALIS.* Ph.D. Dissertation. Karlsruhe Institute of Technology.

[3] Darko Anicic, Paul Fodor, Sebastian Rudolph, and Nenad Stojanovic. 2011. EP-SPARQL: a unified language for event processing and stream reasoning. In *WWW*. ACM, 635–644.

[4] Davide Francesco Barbieri and Emanuele Della Valle. 2010. A Proposal for Publishing Data Streams as Linked Data - A Position Paper. In *LDOW (CEUR Workshop Proceedings)*, Vol. 628. CEUR-WS.org.

[5] Christoph Böhm, Johannes Lorey, and Felix Naumann. 2011. Creating voiD descriptions for Web-scale data. *J. Web Sem.* 9, 3 (2011), 339–345.

[6] Pieter Bonte, Riccardo Tommasini, Emanuele Della Valle, Filip De Turck, and Femke Ongenae. 2018. Streaming MASSIF: Cascading Reasoning for Efficient Processing of IoT Data Streams. *Sensors* 18, 11 (2018), 3832.

[7] François Bry and Michael Eckert. 2007. Twelve Theses on Reactive Rules for the Web. In *Event Processing (Dagstuhl Seminar Proceedings)*, Vol. 07191.

[8] Jean-Paul Calbimonte. 2014. RDF Stream Processing: Let's React. In *OrdRing@ISWC (CEUR Workshop Proceedings)*, Vol. 1303. CEUR-WS.org, 1–10.

[9] Jean-Paul Calbimonte and Óscar Corcho. 2014. Evaluating SPARQL Queries over Linked Data Streams. In *Linked Data Management*. Chapman and Hall/CRC.

[10] Jeremy J. Carroll, Christian Bizer, Patrick J. Hayes, and Patrick Stickler. 2005. Named graphs. *J. Web Sem.* 3, 4 (2005), 247–267.

[11] Gianpaolo Cugola and Alessandro Margara. 2012. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv.* 44, 3 (2012), 15:1–15:62.

[12] C. J. Date. 1984. Some Principles of Good Language Design (with especial reference to the design of database languages). *SIGMOD Record* 14, 3 (1984), 1–7.

[13] Daniele Dell'Aglio, Minh Dao-Tran, Jean-Paul Calbimonte, Danh Le Phuoc, and Emanuele Della Valle. 2016. A Query Model to Capture Event Pattern Matching in RDF Stream Processing Query Languages. In *EKAW (Lecture Notes in Computer Science)*, Vol. 10024. 145–162.

[14] Daniele Dell'Aglio, Emanuele Della Valle, Frank van Harmelen, and Abraham Bernstein. 2017. Stream reasoning: A survey and outlook. *Data Science* Preprint (2017).

[15] Daniele Dell'Aglio and Emanuele Della Valle. 2014. Incremental Reasoning on RDF Streams. In *Linked Data Management*. Chapman and Hall/CRC, 413–435.

[16] Daniele Dell'Aglio, Emanuele Della Valle, Jean-Paul Calbimonte, and Óscar Corcho. 2014. RSP-QL Semantics: A Unifying Query Model to Explain Heterogeneity of RDF Stream Processing Systems. *Int. J. Semantic Web Inf. Syst.* 10, 4 (2014).

[17] Stefano Germano, Thu-Le Pham, and Alessandra Mileo. 2015. Web Stream Reasoning in Practice: On the Expressivity vs. Scalability Tradeoff. In *RR (Lecture Notes in Computer Science)*, Vol. 9209. Springer, 105–112.

[18] Olaf Görlitz and Steffen Staab. 2011. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. In *COLD (CEUR Workshop Proceedings)*, Vol. 782.

[19] Jim Gray (Ed.). 1993. *The Benchmark Handbook for Database and Transaction Systems (2nd Edition).* Morgan Kaufmann.

[20] Thomas R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum.-Comput. Stud.* 43, 5-6 (1995), 907–928.

[21] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. 2013. Observing Linked Data Dynamics. In *ESWC (Lecture Notes in Computer Science)*, Vol. 7882. Springer, 213–227.

[22] Robin Keskisärkkä. 2016. Representing RDF Stream Processing Queries in RSP-SPIN. In *International Semantic Web Conference (Posters & Demos) (CEUR Workshop Proceedings)*, Vol. 1690. CEUR-WS.org.

[23] Alessandro Margara, Jacopo Urbani, Frank van Harmelen, and Henri E. Bal. 2014. Streaming the Web: Reasoning over dynamic data. *J. Web Sem.* 25 (2014), 24–44.

[24] Michael Mathioudakis and Nick Koudas. 2010. TwitterMonitor: trend detection over the twitter stream. In *SIGMOD Conference*. ACM, 1155–1158.

[25] Yves Raimond and Samer Abdallah. 2007. The event ontology.

[26] Mikko Rinne, Eva Blomqvist, Robin Keskisärkkä, and Esko Nuutila. 2013. Event Processing in RDF. In *WOP (CEUR Workshop Proceedings)*, Vol. 1188.

[27] Thomas Scharrenbach, Jacopo Urbani, Alessandro Margara, Emanuele Della Valle, and Abraham Bernstein. 2013. Seven Commandments for Benchmarking Semantic Flow Processing Systems. In *ESWC (Lecture Notes in Computer Science)*, Vol. 7882. Springer, 305–319.

[28] Yehia Abo Sedira, Riccardo Tommasini, and Emanuele Della Valle. 2017. Towards VoIS: A Vocabulary of Interlinked Streams. In *DeSemWeb@ISWC (CEUR Workshop Proceedings)*, Vol. 1934. CEUR-WS.org.

[29] Juan F. Sequeda and Óscar Corcho. 2009. Linked Stream Data: A Position Paper. In *SSN (CEUR Workshop Proceedings)*, Vol. 522. CEUR-WS.org, 148–157.

[30] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. 2009. LODE: Linking Open Descriptions of Events. In *ASWC (Lecture Notes in Computer Science)*, Vol. 5926. Springer, 153–167.

[31] Michael Stonebraker, Ugur Çetintemel, and Stanley B. Zdonik. 2005. The 8 requirements of real-time stream processing. *SIGMOD Record* 34, 4 (2005), 42–47.

[32] Heiner Stuckenschmidt, Stefano Ceri, Emanuele Della Valle, and Frank van Harmelen. 2010. Towards Expressive Stream Reasoning. In *Semantic Challenges in Sensor Networks (Dagstuhl Seminar Proceedings)*, Vol. 10042. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany.

[33] Kia Teymourian, Malte Rohde, and Adrian Paschke. 2012. Knowledge-based processing of complex stock market events. In *EDBT*. ACM, 594–597.

[34] Riccardo Tommasini, Pieter Bonte, Emanuele Della Valle, Erik Mannens, Filip De Turck, and Femke Ongenae. 2016. Towards Ontology-Based Event Processing. In *OWLED (Lecture Notes in Computer Science)*, Vol. 10161. Springer, 115–127.

[35] Riccardo Tommasini, Pieter Bonte, Emanuele Della Valle, Femke Ongenae, and Filip De Turck. 2018. A Query Model for Ontology-Based Event Processing over RDF Streams. In *EKAW (Lecture Notes in Computer Science)*, Vol. 11313. Springer.

[36] Riccardo Tommasini, Yehia Abo Sedira, Daniele Dell'Aglio, Marco Balduini, Muhammad Intizar Ali, Danh Le Phuoc, Emanuele Della Valle, and Jean-Paul Calbimonte. 2018. VoCaLS: Vocabulary and Catalog of Linked Streams. In *International Semantic Web Conference (2) (Lecture Notes in Computer Science)*, Vol. 11137. Springer, 256–272.

[37] Riccardo Tommasini and Emanuele Della Valle. 2017. Challenges & Opportunities of RSP-QL Implementations. In *WSP/WOMoCoE@ISWC (CEUR Workshop Proceedings)*, Vol. 1936. CEUR-WS.org, 48–57.

[38] Riccardo Tommasini and Emanuele Della Valle. 2017. Yasper 1.0: Towards an RSP-QL Engine. In *International Semantic Web Conference (Posters, Demos & Industry Tracks) (CEUR Workshop Proceedings)*, Vol. 1963. CEUR-WS.org.

[39] Riccardo Tommasini, Emanuele Della Valle, Marco Balduini, and Daniele Dell'Aglio. 2016. Heaven: A Framework for Systematic Comparative Research Approach for RSP Engines. In *ESWC (Lecture Notes in Computer Science)*, Vol. 9678. Springer, 250–265.

[40] Riccardo Tommasini, Emanuele Della Valle, Andrea Mauri, and Marco Brambilla. 2017. RSPLab: RDF Stream Processing Benchmarking Made Easy. In *International Semantic Web Conference (2) (Lecture Notes in Computer Science)*, Vol. 10588. Springer, 202–209.

[41] Emanuele Della Valle, Stefano Ceri, Frank van Harmelen, and Dieter Fensel. 2009. It's a Streaming World! Reasoning upon Rapidly Changing Information. *IEEE Intelligent Systems* 24, 6 (2009), 83–89.

[42] Emanuele Della Valle, Daniele Dell'Aglio, and Alessandro Margara. 2016. Taming velocity and variety simultaneously in big data with stream reasoning: tutorial. In *DEBS*. ACM, 394–401.

[43] Emanuele Della Valle, Riccardo Tommasini, and Marco Balduini. 2018. Engineering of Web Stream Processing Applications. In *Reasoning Web (Lecture Notes in Computer Science)*, Vol. 11078. Springer, 223–226.

[44] Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *J. Web Sem.* 9, 2 (2011), 128–136.

[45] Roel Wieringa. 2014. *Design Science Methodology for Information Systems and Software Engineering.* Springer.

Robust Detection of Cyberbullying in Social Media

Mengfan Yao
University at Albany
Department of Computer Science
myao@albany.edu

ABSTRACT

The potentially detrimental effects of cyberbullying have led to the development of numerous automated, data–driven approaches, with an emphasis on classification accuracy. Cyberbullying, as a form of abusive online behavior, although not well–defined, is a repetitive process, i.e., a sequence of aggressive messages sent from a bully to a victim over a period of time with the intent to harm the victim. Existing work has focused on aggression (i.e., using profanity to classify toxic comments independently) as an indicator of cyberbullying, disregarding the repetitive nature of this harassing process. However, raising a cyberbullying alert immediately after an aggressive comment is detected can lead to a high number of false positives. At the same time, three key practical challenges remain unaddressed: (i) detection timeliness, which is necessary to support victims as early as possible, (ii) scalability to the staggering rates at which content is generated in online social networks, (iii) reliance on high quality annotations from human experts for training of highly accurate supervised classifiers.

To overcome the challenges associated with cyberbullying detection in online social networks, my PhD thesis focuses on a novel formulation of the online classification problem as sequential hypothesis testing that seeks to drastically reduce the number of features used while maintaining high classification accuracy. To reduce the dependency on labeled datasets, I seek to develop efficient semisupervised methods that extrapolate from a small seed set of expert annotations. Preliminary results are very encouraging, showing significant improvements over the state–of–the–art.

CCS CONCEPTS

• **Information systems** → *Collaborative and social computing systems and tools*; *Social networking sites*; *Data mining*; *World Wide Web*; *Social networks*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Classification; cyberharassment; optimization; sequential selection; social networks

ACM Reference Format:
Mengfan Yao. 2019. Robust Detection of Cyberbullying in Social Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3314196

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314196

1 PROBLEM STATEMENT

The potentially devastating real–world consequences to victims have resulted in numerous cyberharassmet classification methods with a focus on detection accuracy (e.g. [6]. While high accuracy is undoubted, the state–of–the–art relies on a **fixed** set of features learned during training for **offline** detection (i.e., after all correspondence has become available), hindering the ability to respond in a **timely** manner (i.e., as soon as possible) to cyberbullying events. Moreover, the **scalability** of existing methods to the staggering rates at which content is generated (e.g., 95 million photos and videos are shared on Instagram per day[1]) remains unaddressed [8].

To achieve timely, scalable, and accurate detection of cyberbullying, in my PhD thesis, I would like to address the following 3 research **problems**: (RQ1) Can the time to detection be reduced given that cyberbullying is a repetitive process that unfolds over time? (RQ2) Is it possible to maintain high classification accuracy (i.e., minimize false negatives) while solving for problem 1? (RQ3) Can novel unsupervised learning methods be devised to achieve the same goal?

2 STATE OF THE ART

State-of-the-art mainly comprises 3 bodies of research: (i) cyberbullying detection, (ii) online streaming feature selection (OSFS) methods and (iii) online learning algorithms for classification.

Cyberbullying Detection. Although many approaches have been proposed recently to detect cyberbullying (e.g. [14]), all existing works focus on classification accuracy, neglecting the equally important aspects of scalability and timeliness [8]. With the exception of an incremental cyberbullying detection method to improve detection responsiveness [8], no prior or related work has studied cyberbullying as a repetitive process.

Online Classification Methods. Machine learning methods are usually trained offline and deployed without further updates once training is complete. Once deployed, the accuracy of offline–trained models often deteriorates with time, whereas their retraining may be prohibitive if data is large and/or evolving. At the same time, online learning algorithms that examine data points one at a time, updating their model parameters as new samples arrive, are facing scalability constraints [5]. Although related, my framework fundamentally differs from online classification methods in that the belief on a classification outcome is updated at each time step, allowing detection in real–time.

Online Streaming Feature Selection Methods. Recently, online streaming feature selection (OSFS) methods have been proposed to handle streaming features (e.g.[11]). In contrast to conventional feature selection methods, which are often employed to identify a

[1] 33 Mind–Boggling Instagram Stats & Facts for 2018: https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

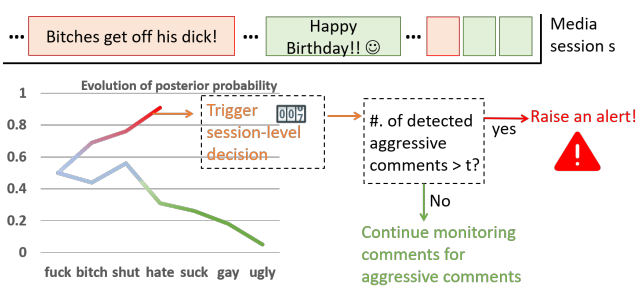


Figure 1: Overview of my proposed approach. Given a media sessions s, and an alert threshold t, my approach examines comments as they become available over time and raises an alert only after the number of detected aggressive comments surpasses the threshold. The posterior probability evolution of an aggressive (red) and non–aggressive (green) comment as more features are examined is provided for illustration purposes. Notice that the number of features used to make a decision in each case differs.

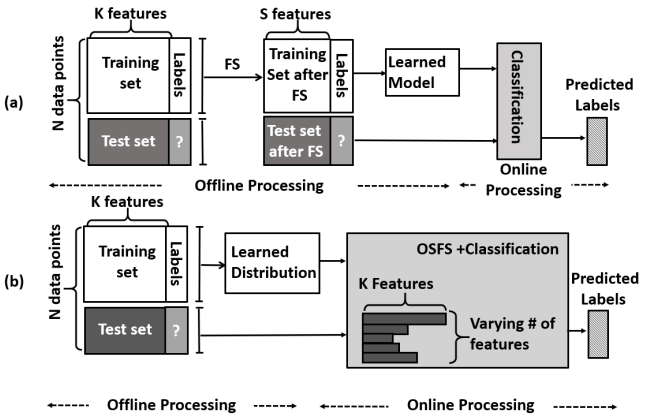


Figure 2: Given a $N \times K$ data matrix and column vector Y of labels, feature selection (FS) is applied to select $S << K$ features before training and classification in (a) existing classification methods. Compare this with (b) my framework, where a model is trained offline once on all features, and online streaming feature selection (OSFS) and classification are performed simultaneously, resulting in a variable number of features used for each data point.

small subset of imformative features to improve classification accuracy but require all features to be available upfront, OSFS methods assume that the size of training data is fixed, and strive to maintain a feature subset that is sequentially updated with the arrival of each new feature based on relevancy and redundancy [11, 15] and group membership [7, 13].

3 PROPOSED APPROACH

In my completed work, I have formulated cyberbullying detection on Instagram media sessions as a two–stage, online framework (illustrated in Figure 1) that (i) sequentially examines comments as they become available over time, and (ii) minimizes the number of feature evaluations necessary for a decision to be made for each comment. This work has been accepted as a short paper to be included in the proceedings of TheWebConf'19 [12].

In my formulation, I use a general data representation, applicable to a wide variety of social media platforms as follows. Each media session $s \in \mathcal{M}$ belongs to a user $u \in \mathcal{U}$, has an associated media object (i.e., image or video) along with its corresponding caption and hashtags, and a set of comments $\{m_1, \ldots, m_{N_s}\}$ from users in $\mathcal{U}$, where N_s denotes the number of comments in s.

In the first stage, comments in s are examined and classified sequentially as aggressive or non–aggressive. Whenever an aggressive comment is detected, a comment–level decision is made. One novelty aspect of my proposed approach is that in this stage, when a given comment is being examined, unlike any existing feature selection method that outputs a global subset of discriminatory features to be used for classification, my approach conducts online streaming feature selection and comment–level classification simultaneously, resulting in a potentially different number of features used to classification and a potential higher classification accuracy. Figure 2 illustrates this difference between existing classification methods (Figure 2. a) and my framework (Figure 2. b).

In the second stage, before a new comment is examined, the number of comment–level detections is compared to a predefined threshold t, and the media session is classified as cyberbullying when the threshold is exceeded. This is another novelty of my proposed approach – comment level decisions are combined to reduce the number of false alarms.

In order to address RQ2, my plan is to introduce a cost into the optimization function to account for the number of comments examined. Such addition will enable my framework to optimally stop examining comments when a decision can be "safely" reached based on accumulated knowledge. As a result, false negative is expected to be largely reduced. Because of the inherent difficulty of RQ3, currently, the work for this task is in an early stage. Specifically, the problem has been formulated, and the first related works have been identified. Possible solutions are yet to be determined.

4 METHODOLOGY

In this section, I introduce the methodology of my proposed optimization framework.

4.1 Problem Formulation

In my hypothesis testing formulation, only two hypotheses exist: (i) $H_{\mathcal{B}}$, which denotes the true hypothesis that m is an aggressive comment, and (ii) $H_{\mathcal{N}}$, which represents the case where m is a non–aggressive comment. Each comment m is described by a vector of features $f(m) = \{y_1, y_2, \ldots, y_K\}$, where K is the total number of features, and $y_k \in \mathcal{Y}$. For each feature y_n, the probability $p(y_n|H_{\mathcal{B}})$ (similarly $p(y_n|H_{\mathcal{N}})$) of the evaluation of the nth feature to observe value y_n when the true hypothesis is $H_{\mathcal{B}}$ (similarly when the true

hypothesis is $H_{\mathcal{N}}$) is empirically computed. Similarly, the *a priori* probability $P(H_{\mathcal{B}}) = p$ of m being an aggressive comment is also estimated empirically. The probability of m being a non–aggressive comment can be computed as $P(H_{\mathcal{N}}) = 1 - p$.

To calculate the belief for m, the framework evaluates features sequentially as illustrated in Figure 1. When examining each comment, at each step, the framework has to select between stopping and continuing the evaluation process based on the accumulated information thus far and the cost of reviewing additional features. The cost coefficient $c_n > 0$, where $n = 1, \ldots, K$ represents the value of time and effort spent evaluating the nth feature. I additionally consider misclassification costs $C_{ij} \geq 0, i = \mathcal{B}, \mathcal{N}, j = 1, \ldots, L$, where C_{ij} denotes the cost of selecting possibility j when the true hypothesis is H_i, and L denotes the number of decision choices (e.g., aggressive or non–aggressive). I factor misclassification costs into my approach to quantify the relative importance of detection errors.

I now formally describe my proposed sequential evaluation process to minimize the number of features used to accurately classify each comment m. Specifically, my proposed sequential evaluation process comprises a pair (R, D_R) of random variables. Random variable R takes values in the set $\{0, \ldots, K\}$, and indicates the feature that the framework stops at, and 0 indicates that no features were evaluated. Hence it is referred to as *stopping time* in decision theory. Random variable D_R denotes the possibility to select among L possible choices. It depends on R and takes values in the set $\{1, \ldots, L\}$. As an example, consider a case where $L = 3$. In this context, $D_R = 1$ corresponds to "aggressive comment", $D_R = 2$ denotes "non–aggressive comment", and $D_R = 3$ indicates "human expert inspection required". Assuming that the random variables y_n are *independent under each hypothesis* $H_i, i = \{\mathcal{B}, \mathcal{N}\}$, the conditional joint probability of $\{y_1, \ldots, y_n\}$ is given as $P(y_1, \ldots, y_n|H_i) = \prod_{k=1}^{n} p(y_k|H_i), i = \mathcal{B}, \mathcal{N}$. Both the decision to stop at stage n (i.e., the event $\{R = n\}$), and the selection of possibility j (i.e., $\{D_R = j\}$) depend only on the accumulated information $\{y_1, \ldots, y_R\}$ by the stopping time R. Equivalently, features that may be examined in the future are not used.

4.2 Optimization Setup

My goal is to achieve high accuracy using the least number of features for detecting aggression at the comment–level. To minimize the number of features considered, the stopping time R and the classification rule D_R have to be selected. To this end, I first define the following cost function:

$$J(R, D_R) = \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + \sum_{j=1}^{L} \sum_{i=\mathcal{B},\mathcal{N}} C_{ij} P(D_R = j, H_i) \right\}. \quad (1)$$

The first expression in the cost function regularizes the number of features, whereas the second expression, commonly referred to as Bayes Risk, penalizes the average cost of the classification rule. My goal can be interpreted as finding the minimum average cost with respect to both random variables R and D_R, i.e., $\min_{R, D_R} J(R, D_R)$, to derive the optimal stopping and classification rules. Intuitively, the optimal rule is to stop at corresponding stopping time R using optimum classification rule D_R.

4.3 Classification Rule

If denote *a posteriori probability* as $\pi_n \triangleq P(H_{\mathcal{B}}|y_1, \ldots, y_R)$, it is not difficult to see that π_n can be iteratively computed as

$$\pi_n = \frac{p(y_n|H_{\mathcal{B}})\pi_{n-1}}{\pi_{n-1}p(y_n|H_{\mathcal{B}}) + (1 - \pi_{n-1})p(y_n|H_{\mathcal{N}})}, \quad (2)$$

where π_{n-1} denotes the posterior probability of a comment being a cyberbullying comment given the first $n-1$ features, and $\pi_0 = p$, i.e., *a priori* probability. Consequently, Eq. (1) can be written compactly as:

$$J(R, D_R) = \mathbb{E}\left\{ \sum_{n=1}^{R} c_n \right\} + \mathbb{E}\left\{ \sum_{j=1}^{L} (C_{Bj}\pi_R + C_{Nj}(1 - \pi_R)) \mathbb{1}_{\{D_R = j\}} \right\}. \quad (3)$$

In order to obtain the optimal classification rule D_R for any stopping time R (i.e., find the optimum Bayes test given that the values of the first R features $y_1, \ldots, y_R$ are observed), an independent of D_R lower bound for the second part of Eq. (3) is needed. Since D_R contributes only to this portion of the average cost, the optimal classification rule D_R for a given stopping time R can then be derived. Specifically, for any classification rule D_R given stopping time R, $\sum_{j=1}^{L} (C_{Bj}\pi_R + C_{Nj}(1 - \pi_R)) \mathbb{1}_{\{D_R = j\}} \geq g(\pi_R)$, where $g(\pi_R) \triangleq \min_{1 \leqslant j \leqslant L} \left[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1 - \pi_R) \right]$. The optimal rule is thus defined as follows:

$$\mathcal{D}_R^{optimal} = \arg\min_{1 \leqslant j \leqslant L} \left[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1 - \pi_R) \right]. \quad (4)$$

From Eq. (4), it follows that $J(R, \mathcal{D}_R^{optimal}) \leq J(R, D_R)$ since the optimal classification rule results to the smallest average cost. Based on the last observation, Eq. (3) can be written as:

$$\min \tilde{J} \triangleq \min_{D_R} J(R, D_R) = \min \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + g(\pi_R) \right\}, \quad (5)$$

which depends only on the stopping time R.

4.4 Stopping Rule

The solution for optimizing $\tilde{J}$ in Eq. (5) with respect to R can be determined by solving the optimization problem $\min_{R \geq 0} \tilde{J}_R = \min_{R \geq 0} \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + g(\pi_R) \right\}$. This constitutes a classical problem in optimal stopping theory for Markov processes [9]. Since the stopping time R can take values in $\{0, 1, \ldots, K\}$, the optimum strategy will consist of a maximum of $K + 1$ stages. In addition, Bellman's principle of optimality [1] states that the solution we seek must also be optimum, if instead of the first stage we start from any intermediate stage and continue toward the final stage. We derive the optimal stopping rule, $R^{optimal}$, as follows:

$$R^{optimal} = \min \left\{ 0 \leq n \leq K | S_n = \tilde{J}_n \right\}, \quad (6)$$

where for $n = K - 1, \ldots, 0$, the n–th stage cost $S_n(\pi_n)$ is related to $S_{n+1}(\pi_{n+1})$ through the following recursion:

$$S_n(\pi_n) = \min \left[g(\pi_n), c_n + \int A_n(y_{n+1}) \times S_{n+1}\left(\frac{p(y_{n+1}|H_{\mathcal{B}})\pi_n}{A_n(y_{n+1})} \right) dy_{n+1} \right], \quad (7)$$

$A_n(y_{n+1}) \triangleq \pi_n p(y_{n+1}|H_{\mathcal{B}}) + (1-\pi_n)p(y_{n+1}|H_{\mathcal{N}})$ and $S_K(\pi_K) = g(\pi_K)$.

The optimal stopping rule derived by Eq. (7) has a very intuitive structure. *i.e.,* stop at the stage where the cost of stopping (the first expression in the minimization) is no greater than the expected cost of continuing given all information accumulated at the current stage (the second expression in the minimization). Specifically, at each stage n, my method faces two options given π_n: (i) stop evaluating features and classify, i.e. compute D_n, or (ii) continue and evaluate the next feature. Note that $\tilde{J}_n$ can be interpreted as the cost if the system stops at stage n, whereas S_n is the optimal cost when the system is at stage n (regardless of continuing or stopping). From Eq. (7), it is easy to verify that $S_n \geq \tilde{J}_n$ for all $n = 0, \ldots, K$. Thus, the optimal time to stop is when the equality is first obtained.

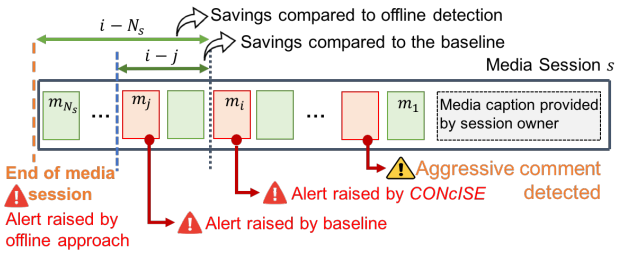


Figure 3: Illustration of timeliness, i.e., the number of comments "saved" as compared to (i) an offline approach that has to wait until all comments related to a media session become available, or (ii) an online baseline that raises an alert after a threshold.

5 RESULTS

In this section I provide experimental analysis to evaluate my proposed approach (termed CONcISE). For a fair comparison with the state–of–the–art, I use the Instagram data set collected by Hosseinmardi et al. [6]. In total, the dataset contains 3,829,756 users and 9,828,760 comments. Of all media sessions which contain at least 40% profanities. I augmented this dataset with $\sim 10K$ comment–level labels obtained from 10 experts. The comments span in total 22.1% of the media sessions with 40% or more profanity. I use this small subset of labeled comments for training, and the remaining unlabeled comments (i.e., 77.9% of labeled media sessions in the original datasetf from [6]) for testing.

As for feature representations, I consider the 10 unigrams [6], the 1,384 profane unigrams and bigrams [10] and 394 offensive unigrams and bigrams [4] that are shown to be imformative in state-of-the-art approaches [6, 8], to examine the robustness of *CONcISE* on the dictionary used for detection. This results in three variants of *CONcISE*: **CONcISE–10**, **CONcISE–Profane**, and **CONcISE–Noswearing**, respectively, that use the corresponding dictionaries.)

Comparison with Cyberbullying Detection State-of-the-art. Table 1 summarizes the performance of *CONcISE* as compared to the state–of–the–art for cyberbullying detection, in terms of accuracy, recall and precision of the bullying class, area under the ROC curve (AUC), the average number (and standard deviation) of features used by my approach in detecting aggressive comments (similarly for the number of comments needed to detect a cyberbullying session). The best performing method is marked in bold. All in all, *CONcISE* and its variations outperform the baselines almost invariably across the board of evaluation metrics. Clearly, all three variants of *CONcISE* outperform the baselines, often by a considerable margin. At the same time, *CONcISE* significantly reduces both the average number of features used to classify individual comments as well as the number of comments examined before determining if a session is an instance of cyberbullying.

Comparison with Online Streaming Feature Selection State-of-the-art. I compare *CONcISE* to the best performing OFSF methods in terms of feature selection quality, namely SAOLA [13] and OFS–Density [15], over the feature space defined by the corresponding disctionary used, i.e., 10 unigrams, Noswearing and Profane dictionaries respectively. As SAOLA and OFS-Density do not perform classification in conjuction to feature selection, I use KNN, SVM, Random Forest and Standard Bayes for both baselines; KNN and SVM have previously been shown to achieve the highest classification performance in [13, 15]. I report the accuracy of *CONcISE*, as well as all combinations of baselines and classifiers. Figure 4 shows that *CONcISE* consistently outperforms OFS-Density and SAOLA over all 3 feature spaces.

Timeliness Analysis. To compare the best performing variance of *CONcISE*, namely, *CONcISE–Noswearing*, against baselines with respect to *timeliness*, I introduce a measure "# of saved comments", defined as $i - j$ where i and j, respectively, denotes the number of comments that *CONcISE* and a given baseline have to inspect before raising an alert. My definition of timeliness along with the above mentioned scenarios are illustrated in Figure 3. The result has shown that *CONcISE–Noswearing* can detect cyberbullying media sessions faster than the best performing offline approach (i.e. RDFS), for all thresholds considered (in addition to being more accurate). As comparing to the best performing online approach (i.e. RF–comment), the result suggests that RF–comment achieves slightly faster detection than *CONcISE–Noswearing*. However, with less than 70% precision (see Table 1), RF–comment tends to classify non–aggressive comments as aggressive. In turn, RF–comment reaches the threshold faster than *CONcISE–Noswearing* in true cyberbullying sessions. The downside to this outcome is a higher number of false positives in sessions that are not in reality cyberbullying sessions.

Scalability Analysis. I evaluate the scalability of *CONcISE* in both classifying comments and in raising a session–level alert. In my non optimized implementation, *CONcISE-10*, *CONcISE-Profane* and *CONcISE-Noswearing* classify on average 45,140 comments, 33,964 comments, and 30,849 comments per second, accordingly. With respect to runtime for session–level decisions, I found baselines RDFS, RF, and TM to be the fastest among all methods, for an average of 0.0021, 0.0011, and 0.002 seconds respectively in making a session level decision. However, such "speed" is meaningless if considered in isolation to timeliness. Both RDFS, RF, and TM require all comments of a media session to be available for classification, making them offline for all practical purposes. The rest of the methods achieve similar runtime (s), ranging from 0.0077 to 0.0153 second per media session, with *CONcISE–10* and RF–comment being the fastest and the slowest, respectively. Even though the difference may seem negligible for this dataset, considering a real–world

Table 1: Performance comparison of *CONcISE* with state–of–the–art cyberbullying detection methods.

Method	Accuracy	Recall	Precision	AUC	Avg. # of features (std.)	Avg. # of comments (std.)
RDFS [6]	0.751	0.449	**0.858**	**0.877**	10 (0)	83
RF [2]	0.791	0.703	0.749	0.862	13 (0)	83
RDFS–comment [6]	0.733	0.463	0.760	0.684	10 (0)	56.15 (36.64)
RF–comment [2]	0.776	0.751	0.699	0.724	13 (0)	33.07 (22.59)
DLR [8]	0.497	0.651	0.616	0.521	4 (0)	35.90 (45.43)
TM [3]	0.749	0.783	0.649	0.816	115(0)	83(0)
CONcISE–10	0.783	**0.794**	0.695	0.864	**2.97 (1.17)**	**26.68 (16.92)**
CONcISE–Profane	**0.806**	0.769	0.745	0.860	3.76 (3.19)	30.62 (21.82)
CONcISE–Noswearing	**0.806**	0.776	0.742	0.862	3.92 (2.16)	29.75 (21.09)

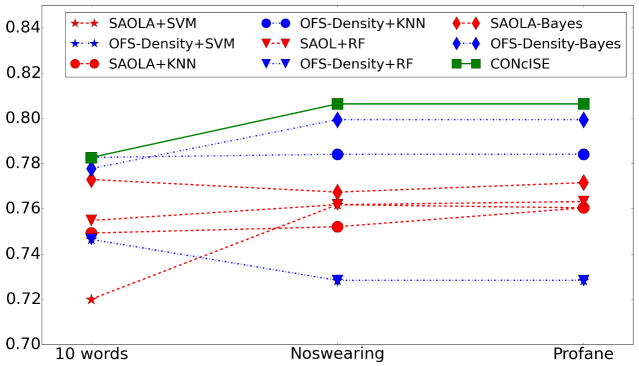


Figure 4: Accuracy comparison of *CONcISE*, and SAOLA and OFS-Density.

scenario where a million media sessions are to be evaluated in real–time. By extrapolating the results of my analysis to such scenario, we can expect *CONcISE–10* to classify all sessions within $\sim 7,600$ seconds (i.e., ~ 2 hours) less than RF–comment.

Finally, note that even though such numbers refer to a sequential implementation of *CONcISE*, my approach can be trivially parallelized by classifying each comment/session in parallel. This empirical result, in addition to the fact that the runtime of *CONcISE* is linearly proportional to the number of features it uses to reach a classification decision, provide strong evidence about the real–world applicability of *CONcISE*.

6 ACKNOWLEDGMENTS

I would like to thank my advisor, Dr.Charalampos Chelmis, for all his help and guidance that he has provided me over the past two years.

7 CONCLUSION AND FUTURE WORK

The main goal of my PhD study is to introduce socio–technical advances that enhance Web platforms. Specifically, I strive to achieve timely and accurate detection of cyberbullying instances in online social networks while being scalable to the staggering rates at which content is generated in online communities. The main research problems of my Ph.D. are (1) Timely and Scalable detection (2) Accurate detection with minimal false negative (3) Unsupervised learning method that achieves the same goals as in (1) and (2).

My preliminary experiments showed the feasibility of the proposed framework, with up to 62.17% of improvement in accuracy and 67.86% reduction in terms of timeliness comparing to baseline approaches.

In future work, I plan to explore features such as user– and network–information and the sequence of conversations to further improve classification accuracy. I am also planning to evaluate the performance of my approach on additional datasets from diverse platforms including Ask.fm and Twitter, which are reported to be key social networking venues where users frequently become victims of cyberbullying. Given the broad definition of cyberbullying, I may also design detection strategies grounded in a more nuanced, multi–dimensional representation of repetitive harassment instead of striving for a global and/or simplified indicator of cyberbullying. The most interesting part of the work will be in finding out how to make use of unsupervised methods, since both cost function and the learning process can be inherently different from the proposed sequential framework which relies on training labels.

REFERENCES

[1] D. P. Bertsekas. 2005. *Dynamic Programming and Optimal Control*. Vol. 1. Athena Scientific.
[2] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 13–22.
[3] Vivek Singh Devin Soni. [n. d.]. Time Reveals AllWounds: Modeling Temporal Dynamics of Cyberbullying Sessions. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*.
[4] AllSlang family. [n. d.]. Internet Slang Swear Word List & Curse Filter. https://www.noswearing.com/dictionary.
[5] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2018. Online Learning: A Comprehensive Survey. *arXiv preprint arXiv:1802.02871* (2018).
[6] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 186–192.
[7] Haiguang Li, Xindong Wu, Zhao Li, and Wei Ding. 2013. Group feature selection with streaming features. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1109–1114.
[8] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. 2018. Scalable and Timely Detection of Cyberbullying in Online Social Networks. (2018).
[9] Albert N Shiryaev. 2007. *Optimal Stopping Rules*. Vol. 8. Springer Science & Business Media.
[10] Luis von Ahn. [n. d.]. Offensive/Profane Word List. https://www.cs.cmu.edu/~biglou/resources/bad-words.txt.
[11] Xindong Wu, Kui Yu, Hao Wang, and Wei Ding. 2010. Online streaming feature selection. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. Citeseer, 1159–1166.
[12] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media. In *2019 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.
[13] Kui Yu, Xindong Wu, Wei Ding, and Jian Pei. 2016. Scalable and accurate online feature selection for big data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 2 (2016), 16.

[14] Rui Zhao and Kezhi Mao. 2017. Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing* 8, 3 (2017), 328–339.

[15] Peng Zhou, Xuegang Hu, Peipei Li, and Xindong Wu. 2019. OFS-Density: A novel online streaming feature selection method. *Pattern Recognition* 86 (2019), 48–61.

Attention workshop Chairs' Welcome

It is our great pleasure to welcome you to the first multi-disciplinary workshop on "Attention - from Neuroscience to the Web and Wellbeing", at the *Web Conference 2019*. In an age where multiple apps, advertisements and an abundance of social media are vying to get the user's attention, attention has become an extremely scarce resource. Never before has attention economy been more acute. Designing effective solutions that are mindful of user's attention call for better understanding of attention -- how it works in the brain, how it can be measured at scale on the web, and its current applications as well as challenges.

To raise the community's awareness around this timely issue, this workshop has been designed with the vision of bringing together leaders in academia and industry to share the state-of-the-art on Attention from diverse disciplines, including its successful applications. We believe this workshop will spearhead new research initiatives towards better use of attention for digital wellbeing on the web and apps.

The workshop includes 2 keynote speakers. Dr. Thales Teixeira from Harvard Business School to speak on the "Economics of Attention", and Dr. Tom Insel from MindStrong Health (ex-director of National Institute of Mental Health) to speak on "Digital Wellbeing".

The workshop will begin with experts showcasing insights from 4+ decades of work in Neuroscience / Cognitive Psychology on the science of attention. This will be followed by talks on different methods to study attention at scale on the web. The afternoon session will focus on the applications of attention in different domains, including Search, Advertising, Design, Brain health and Digital Wellbeing.

We hope that you will find the workshop program interesting, providing you with a valuable opportunity to learn and share ideas with other researchers and practitioners from institutions around the world.

Vidhya Navalpakkam
Google AI

Laurent Itti
University of Southern California

Attention Workshop Organization

Workshop Chairs: Vidhya Navalpakkam *(Google AI, USA)*
Laurent Itti *(University of Southern California, USA)*

Speakers Tom Insel (Mindstrong Health, USA)
Thales S. Teixeira (Harvard Business School)
Jeremy Wolfe (Harvard Medical School, USA)
Zoya Bylinskii (Adobe Research, USA)
Jeff Huang (Brown University, USA)
Michael Xuelin Huang (Google, USA)
Eugene Agichtein (Emory University, USA)
Tanzeem Choudhury (Cornell University, USA)
Jamshid Ghajar (Stanford University, USA)
Joaquin Anguera (UCSF, USA)

Scalable Eye Tracking for Mobile Devices

Michael Xuelin Huang
Google
Mountain View, CA, U.S.A., mxhuang@google.com

ABSTRACT

Eye tracking provides an effective solution to users' attention, interest, and engagement. While gaze estimation based on a standard camera can be versatile, it remains challenging to achieve an accurate, robust, and scalable solution on mobile devices. In this talk, I will describe three studies that aim to address these challenges. Specifically, 1) we found that screen reflection on user's cornea can be leveraged for gaze estimation and it considerably improves the practicability of indoor eye tracking. 2) We exploited gaze-hand coordination and applied interaction data for implicit calibration when a user naturally interacts with the computer. This can prevent users from tedious and intrusive calibration in practice. 3) We also proposed to train a multi-device person-specific gaze estimator to accelerate implicit calibration. It adapts the data from different personal devices to learn the shared mapping from user appearance into eye gaze. Taken together, these studies identify indicative eye gaze features, alleviate user calibration effort, and thus pave the way for scalable eye tracking in daily use.

CCS CONCEPTS

• Human-centered computing → Human-computer interaction (HCI)

KEYWORDS

Implicit calibration, Gaze estimation, Gaze-hand alignment

ACM Reference format:

Michael Xuelin Huang. 2019. Scalable Eye Tracking for Mobile Devices. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317094

1 BIO

Michael Xuelin Huang is a research scientist at Google, Mountain View, where he works on advanced human-computer interaction paradigms at the Android Research Group. Prior to joining Google, he was a postdoc at the Perceptual User Interface Group at Max Planck Institute for Informatics, Germany. He holds a Ph.D. at the Hong Kong Polytechnic University, and he received his M.S. and B.S. degrees at Beihang University, China. His research interests include affective computing, human–computer interaction, gaze estimation, and pattern recognition.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317094

Lessons from neuroscience

Laurent Itti
Computer science department
psychology and neuroscience graduate program, University of Southern California, Los Angeles, CA, USA, itti@usc.edu

ABSTRACT

Visual attention and eye movements in primates have been widely shown to be guided by a combination of stimulus-dependent or 'bottom-up' cues, as well as task-dependent or 'top-down' cues. Both the bottom-up and top-down aspects of attention and eye movements have been modeled computationally. Yet, it is not until recent work which I will describe that bottom-up models have been strictly put to the test, predicting significantly above chance the eye movement patterns, functional neuroimaging activation patterns, or most recently neural activity in the superior colliculus of human or monkey participants inspecting complex static or dynamic scenes. In recent developments, models that increasingly attempt to capture top-down aspects have been proposed. In one system which I will describe, neuromorphic algorithms of bottom-up visual attention are employed to predict, in a task-independent manner, which elements in a video scene might more strongly attract attention and gaze. These bottom-up predictions have more recently been combined with top-down predictions, which allowed the system to learn from examples (recorded eye movements and actions of humans engaged in 3D video games, including flight combat, driving, first-person, or running a hot-dog stand that serves hungry customers) how to prioritize particular locations of interest given the task. Pushing deeper into real-time, joint online analysis of video and eye movements using neuromorphic models, we have recently been able to predict future gaze locations and intentions of future actions when a player is engaged in a task. In a similar approach where computational models provide a normative gold standard against a particular individual's gaze behavior, machine learning systems have been demonstrated which can predict, from eye movement recordings during as little as only 5 minutes of watching TV, whether a person has ADHD or other neurological disorders. Together, these studies suggest that it is possible to build fully computational models that coarsely capture some aspects of both bottom-up and top-down visual attention.

CCS CONCEPTS

• Human-centered computing~Interaction design theory, concepts and paradigms

KEYWORDS

Attention, neuroscience, eye movements, neurodevelopmental disorders, computational modeling.

ACM Reference format:

Laurent Itti. 2019. Lessons from neuroscience. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317091

1 BIO

Laurent Itti received his M.S. degree in Image Processing from the Ecole Nationale Superieure des Telecommunications (Paris, France) in 1994, and his Ph.D. in Computation and Neural Systems from Caltech (Pasadena, California) in 2000. He has since then been an Assistant, Associate, and now Full Professor of Computer Science, Psychology, and Neuroscience at the University of Southern California. Dr. Itti's research interests are in biologically-inspired computational vision, in particular in the domains of visual attention, scene understanding, control of eye movements, and surprise. This basic research has technological applications to, among others, video compression, target detection, and robotics. Dr. Itti has co-authored over 150 publications in peer-reviewed journals, books and conferences, three patents, and several open-source neuromorphic vision software toolkits.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05
https://doi.org/10.1145/3308560.3317091

5th AW4City 2019 Chairs' Welcome

It is our great pleasure to welcome you to the *AW4City 2019 5th International Workshop on Web Applications and Smart Cities*. This year's fifth AW4City version follows up the success of the previous four events and emphasizes on the contribution of web applications and Apps to urban mobility. In the era of cities and under the UN 2030 Agenda for sustainable growth, cities are making an important shift regarding thinking of compact cities and they develop sustainable mobility plans. Issues like parking, traffic congestion and safety, sharing transportation and smart parking are among the topics where governments and the industry direct smart services, while data collection and analysis enable government's fast –or even real-time- decision-making for alternative transportation means' deliberation, public transportation changes, urban planning etc., and new business models appear that utilize platforms and data. Additionally, new types of business in transportation appear (e.g., that utilize sensors and big data), while things' transportation is another challenge that cities have been started phasing.

Our call received 15 articles from United States, Europe and Asia, which followed a peer review process regarding their relevance, quality, and novelty and 8 full and 1 short paper were selected. Review and acceptance statistics are as follows:

Submission Type	Reviewed	Accepted	
Full Technical Papers	13	8	61%
Short Papers	2	1	50%

The beginning of this half-day event contains the following three talks:

- Emerging mobility and the web
- Smart city platforms for digital twinning in mobility
- Mobility Applications and their role safer and a more inclusive cities

Then, a panel of presenters covers topics with regard to:

- Artificial intelligence and the Web for urban traffic and mobility management;
- Mobility-as-a-Service in cities;
- Web applications for smart tourism services in cities;
- Citizen participation via the Web for mobility issue handing;

We expect that you will find AW4City's program interesting, which will enable you to learn and share ideas with other researchers and practitioners from institutions around the world.

Leonidas Anthopoulos
Workshop Chair
University of Thessaly, Greece

Marijn Janssen
Workshop Chair
Delft University of Technology, the Netherlands

Vishanth Weerakkody
Workshop Chair
University of Bradford, UK

5th AW4City 2019 Chairs' Welcome

Workshop/Track Chairs: Leonidas Anthopoulos (University of Thessaly, Greece)
Marijn Janssen (Delft University of Technology, The Netherlands)
Vishanth Weerakkody (University of Bradford, UK)

Program Committee: Dirk AHLERS *(Norwegian University of Science and Technology)*
Cristina Maria BUETI *(International Telecommunication Union (ITU))*
Stephen COHEN *(Microsoft, USA)*
Beth COLEMAN *(University of Waterloo, Director of City as Platform)*
Nicolas DOUAY *(MCF Paris 7 / UMR Geographie-Cites / LabEx DynamiTe)*
Ramy Ahmed FATHY *(Senior IEEE)*
Gordon FELLER *(Meeting of the Minds (meetingoftheminds.org))*
Toru ISHIDA *(Department of Social Informatics, Kyoto University, Japan)*
Vinay KANDPAL *(University of Petroleum & Energy Studies (UPES), India)*
Sakano NARUTOSHI *(Fujitsu, Japan)*
Sofia Toufic SHAWYRI *(University of California at Berkeley)*
Alois PAULIN *(Siemens, Austria)*
Arthur RIEL *(The World Bank, Washington, USA)*
Christopher G. REDDICK *(The University of Texas at San Antonio, USA)*
Markus RITTENBRUCH *(Queensland University of Technology, Australia)*
Mag. Michael SACHS *(Donau-Universitat Krems)*
Flora SALIM *(RMIT University, Melbourne, Australia)*
Júlia SEIXAS *(NOVA University of Lisbon, Portugal)*

Reviewers: Amit AGARWAL
H. Michael CHUNG
Rob CHRISTIAANSE
Hinnerk EISSFELDT
Daniel HERZOG
Luis-Daniel IBÁÑEZ
Vinay KANDPAL
Aravinda KUMARAN
Eddy MADDALENA
Babis MAGOUTAS
Md.Ashifuddin MONDAL
William MONEY
Vaia MOUSTAKA
Zeenat REHENA
Daksh SHAH
Nicolas TEMPELMEIER

THE WEB CONFERENCE

Persuasive Technologies for Sustainable Smart Cities: The Case of Urban Mobility

Evangelia Anagnostopoulou[†]
ICCS, National Technical University of Athens, Greece, eanagn@mail.ntua.gr

Babis Magoutas
ICCS, National Technical University of Athens, Greece, elbabmag@mail.ntua.gr

Efthimios Bothos
ICCS, National Technical University of Athens, Greece, mpthim@mail.ntua.gr

Gregoris Mentzas
ICCS, National Technical University of Athens, Greece, gmentzas@mail.ntua.gr

ABSTRACT

In this paper, we study the effectiveness of personalized persuasive interventions to change urban travelers' mobility behavior and nudge them towards more sustainable transport choices. More specifically, we embed a set of persuasive design elements in a route planning application and investigate how they affect users' travel choices. The design elements take into consideration the style, the intensity, the target of persuasive interventions as well as users' characteristics and the trip purpose. Our results show evidence that our proposed approach motivates users on a personal level to change their mobility behavior and make more sustainable choices. Furthermore, by personalizing the persuasive interventions while considering combinations of interventions styles (in our case messages and visualizations) as well as adjusting the intensity of persuasive interventions according to the trip purpose and the transport modes of the routes which the user is nudged to follow, the effects of the persuasive interventions can be increased.

KEYWORDS

Persuasive technology, smart city, sustainability, urban mobility, design elements, personalization

1 Introduction

Humanity's future is set to be urban. Currently, over half of the global population lives in urban areas and 1.5 million people are added to urban environments every week [1]. This rapid population growth results to two conflicting issues. On the one hand, it leads to overexploitation of resources and an inadequate number of services failing to cover existing needs. On the other hand, sustainable development must be achieved to ensure citizens quality of life and well-being. Smart cities that leverage technological developments, including mobile technologies, sensors and connected infrastructures hold the promise of providing solutions to address problems from increased urbanization, including those concerning environmental sustainability and governance of services for efficient use of resources.

In current urban settings, rendering transport systems sustainable is a major problem as such systems have significant impacts on the environment, accounting for about 25% of the world energy consumption and carbon dioxide emissions [2]. Furthermore, recent reports show that the greenhouse gas (GHG) emissions from transport are increasing at a faster rate than any other energy using sector, especially in urban environments [3], rendering current practices unsustainable.

One way to improve the sustainability of mobility in smart cities is by increasing citizens' awareness of the negative impacts of their transport choices and proving means to nudge them towards the use of sustainable transport modes. This can be achieved by designing and implementing approaches based on information and communication technologies (ICT) that make travelers' aware of the environmental impact of travel mode

choices and by providing personalized persuasive interventions towards adopting transportation habits that result to less car usage and more use of public transportation, bicycles and walking.

Our work builds on the concept of persuasive technologies (PT) which guide users towards the adoption of desired attitudes or actions by infusing persuasive strategies in ICT applications that interact with the users [4]. PT have been successfully utilized in several application domains to support sustainable behaviors, such as health, education and environmental sustainability, including transportation (see e.g., [5], [6]).

More specifically, we study the effectiveness of personalized persuasive interventions to change user's transport behaviors towards more sustainable choices. For the purposes of our study we designed and developed a personalized persuasive approach which was integrated in a route planning mobile application. A set of design elements were defined which were tested in terms of their ability to influence the effectiveness of persuasive interventions in three real life pilots where users from the cities of Vienna (Austria), Ljubljana (Slovenia) and Birmingham (UK) used the route planning application for a period of 12 weeks. The outcomes of our study can be used by smart city transport authorities and persuasive systems designers in order to develop effective persuasive technologies for sustainable transportation.

The remainder of the paper is organized as follows. Section 2 provides more details on the concept of persuasive technologies and discusses the related work; Section 3 formulates the design elements of PT for urban transportation which are the focus of this paper; Section 4 provides an overview of the approach and system that was developed by considering the aforementioned design elements, and explains its main assumptions and functionalities; Section 5 presents the results which came out of the real life pilots; Finally, Section 6 concludes the paper with our final remarks and suggestions for future work.

2 Background and Related Work

A number of applications have been developed over the last years which support users in shifting their behaviour towards more sustainable means of transportation by employing persuasive technologies. One of the first mobile persuasive applications was UbiGreen [7], which adapts the background graphics of the phone to provide visual feedback that aims to reduce car use and to encourage greener alternatives, including carpooling, public transport, and pedestrian modalities. Other examples of such applications include PEACOX [8], TRIPZOOM [9] and SUPERHUB [11].

According to the literature review by Anagnostopoulou et al. [12], several persuasive strategies such as behaviour feedback, social comparison, goal-setting, gamification, personalized suggestions and challenges have been used so far in PT for sustainable transportation, and new ones are being continuously developed. Indicative examples of how persuasive strategies are implemented include the following. Behavioural feedback: tracking a user's behaviour and providing visual feedback on the emissions caused by her/his choices; social comparison: comparing one's own mobility behaviour to that of others and presenting comparative information; personalized suggestions: providing system-generated suggestions that urge users to follow more environmental transportation modes. The implementation details (e.g. mobile route planning app versus web-based systems) of PT for sustainable transportation are quite diverse however, findings from the literature review of Klecha and Gianni [13] show that mobile devices are being prevalently used.

Regarding the effectiveness of such applications on behavioural change, a recent review of existing systems designed to promote sustainable travel behaviour [14] concludes that effect sizes are mostly small and methodologically robust studies are largely missing, hence no definitive conclusion yet can be derived.

An important and active area of research in persuasive technologies focuses on their personalization. Existing applications that integrate personalized persuasion try to personalize specific aspects of a single persuasive strategy and not the persuasive strategy per se. For example, in [15] an approach of personalizing challenges (competition strategy) is described, while in [7] an application that persuades users to make more sustainable choices through personalized suggestions and self-monitoring is implemented. In our previous work ([16]), we describe a first attempt to nudge users on a personalized level in order to change their mobility behavior and make more sustainable choices. Although the results were fairly positive, further exploration of persuasive strategies' personalization for behavioural change towards sustainable mobility is required. In more details, the full potential of persuasive technology to change user mobility behaviour remains untapped

due to the fact that the effectiveness of persuasive technology to nudge users towards sustainable mobility choices, depends on several features that have not been considered in a holistic and consistent manner in the design of systems used to deliver the persuasive interventions (e.g. mobile apps with their back-end services). In this paper we focus on the study of a set of features, hereafter called Design Elements (DE), which include the: personalization of persuasive interventions based on the characteristics of the individual (DE1); adaptation of interventions to the results of previous persuasive interactions (DE2); combination of different styles and forms of interventions (DE3); degree of persuasive interventions' intensity (DE4); target of persuasive interventions in terms of the specific transportation mode they are nudging the users to follow (DE5); the adaptation of persuasive interventions' intensity depending on the trip purpose (DE6).

3 Design Elements Influencing Persuasion

The aforementioned DEs along with the way they influence the effectiveness of persuasive interventions, are elaborated in this section. Note that these DEs have been taken into account for the design and development of an approach and the related persuasive system for the delivery of personalized persuasive interventions to change user mobility behaviour. The approach and application are described in Section 4, while Section 5 provides empirical evaluation results which show how the DEs affect the persuasiveness of the application.

DE1: Persuasive interventions specifically addressing each user and tailoring persuasive features to the individual. Personalizing persuasive technology is an important and active area of research at the human computer interaction community. Persuasive technology designers use many different persuasive strategies to change user behaviour, such as behaviour feedback, social comparison, goal-setting, gamification, personalized suggestions and challenges. However, people differ in their susceptibility to different persuasive strategies, i.e. a strategy that may be effective for an individual may not be that effective for another. Personalized approaches can be more successful than "one size fits all" as they can adapt the selected persuasive strategies to specific users, rather than the general audience and can sustain users' interest over time while providing better results [17].

DE2: Incorporation of a learning component that allows the system to adapt the persuasive interventions towards sustainable travel choices based on previous experience. The view that the system has on a particular user, in terms of his/her susceptibility to the different persuasive strategies, can be dynamically and constantly updated through machine learning techniques, as users interact with the system. More specifically, user susceptibility to the different strategies can be dynamically updated by considering previous successful persuasive interactions with that particular user and other similar users. In this way, persuasive interventions can be dynamically adapted to the potentially changing user needs, sustaining their impact over time. The bigger the number of users receiving persuasive interventions and the number of persuasive attempts per user, the faster such a dynamic approach will converge to (i.e. identify) the best user susceptibility to the different persuasive strategies.

DE3: Combining different means of communicating persuasive interventions (e.g. textual messages with graphs) to nudge users to make more sustainable travel choices. Another important factor in the design of persuasive interventions is the way these are delivered. One of the most common methods refers to presenting feedback, suggestions and social comparison in the form of textual messages. Another method refers to presenting feedback about the mobility habits of users. This could be either by e.g. displaying graphs depicting the percentages of transport modes usage, or by showing money spent, calories burnt, and carbon dioxide emissions consumed.

Recent years have witnessed a remarkable increase in the adoption of visualization as a means to convey information through data. Visualizations aid human understanding of data by leveraging the human visual system's highly-tuned ability to see patterns, spot trends, and identify outliers by making data more accessible and appealing [18]. Although visualizations as a communication tool to inform users gains popularity, little research has been done to examine the effects of data visualization in influencing them or in making a message more persuasive [19]. For example in [20] the authors examine social visualisations, i.e. the visualization of individual's data in relation to that of others, as a tool in persuasive technology with the potential to cause behaviour change towards healthier lifestyles through social comparison.

Enabling different means of communicating persuasion, allows us to investigate, whether a message combining narrative and statistical evidence is more persuasive than a message using either narrative (i.e. textual persuasive message) or statistical evidence (i.e. persuasive visualization) alone. Evidence from the literature supports the aforementioned argument; see for example [21].

DE4: Adjusting the intensity of persuasive interventions. The intensity of persuasive interventions is a key factor in the design of persuasive systems as the frequency of persuasive attempts for changing users' behaviour can affect the effectiveness of interventions. For example, energy feedback research [22] suggests that frequent feedback is preferred and is more effective compared to less frequent feedback. It is also possible that frequent feedback can become repetitive to such an extent that users are annoyed [23]. The problem of personalizing the frequency and intensity of persuasive interventions to the needs of an individual recipient has been gaining interest over the last years, with the aim to increase interventions' effectiveness. This is because individuals may differ in how much support they want in general but also around specific moments in behaviour change [24]. For example, when changing behaviours some people may prefer only low frequency contact, others may want more intense support, while still others may need more support under specific situations.

Related studies have provided preliminary results which show that adapting the intensity of the persuasive interventions to the preferences and characteristics of individual users upgrades the interventions' persuasive capabilities [25]. In our previous work we have also confirmed that the intensity of persuasive attempts matters [16]. Therefore, one of the issues to be explored in this work is whether adjusting the intensity of persuasive interventions to the needs of an individual has an impact on the effectiveness of interventions.

DE5: Nudging users to sustain or intensify current green behaviour. A persuasive intervention that nudges users to change their mobility behaviour, includes the definition of the intervention's target in terms of a sustainable transportation mode. It is important to select the target of persuasive interventions based on the current typical mobility choices or habits of the individual, avoiding to driving him out of his/her comfort zone in a big extend. For example, it may be more difficult to nudge individuals who mostly use their car to begin using bike sharing than nudging them to begin combining the use of car with public transportation, i.e. park & ride, in certain occasions. Similarly, an individual who combines car with public transportation is easier to change his habits and take the bus instead of walking for long distances. In other words, since a radical behavioural change is rarely easy, it often requires a gradual progression of small steps toward a larger goal. In the same direction, nudging individuals to sustain or intensify their current green behaviour and supporting them to resist the temptation of using more unsustainable modes of transport seems to be more effective – at least in the short-term – than pushing them to get out of their comfort zone by completely changing their mobility behaviour to much more sustainable modes.

DE6: Targeting persuasive interventions towards environmental friendly modes of transport mainly for leisure than commute trips. A traveller's decision on transport mode selection depends upon the value of travel time savings (VTTSs) [26], a measure used in the domain of transportation to define the value of every minute (time) that individuals save during their travels. The value of VTTSs varies for different trip purposes. For example, if an individual travels (commutes) to her/his work, the VTTSs is high since the individual wants to minimize the travel time as much as possible. Instead, when s/he travels for leisure purposes, the VTTSs is lower than commuting. Especially, when the VTTSs is lower than usual, it is more likely that the individual selects a more environmentally friendly route, which could take longer time to reach a destination (e.g. use of public transportation instead of a car).

Past research [27] has provided evidence on the variation of the VTTSs by country and trip purpose. Specific models are applied to produce VTTSs for business travel, commuting, and for other purposes in passenger transport, for 25 European Union Member states. Based on this study, for all the countries of interest for our work, i.e. Austria, Slovenia and UK, the VTTSs for commuting travels is higher than for leisure travels. Therefore, and after following the findings mentioned above, we assume that the persuasive interventions are more effective when the trip is made for leisure purposes than for commuting.

4 Personalized Persuasion for Sustainable Urban Mobility

For the purposes of our study, we have designed and implemented an approach for personalized persuasion in the context of sustainable mobility that incorporates all the design elements discussed in Section 3. The aim is to support transportation decisions for selecting more environmentally friendly routes that lead to emissions reduction in a smart city context. Our approach is integrated in a route planning mobile application developed as part of the EU-funded H2020 project OPTIMUM[1]. The overall design and architecture of our approach is presented in Figure 1. Our personalized persuasive approach relies on two complementary services: (i) the Route Recommendation Service (RRS) that operates on lists of alternative routes for travelling from A to B provided by a Routing Engine and identifies alternatives that match user context and preferences and are environmentally friendly and (ii) the Personalized Persuasive Service (PPS) that provides personalized persuasive features that nudge users to take an environmentally friendly route.

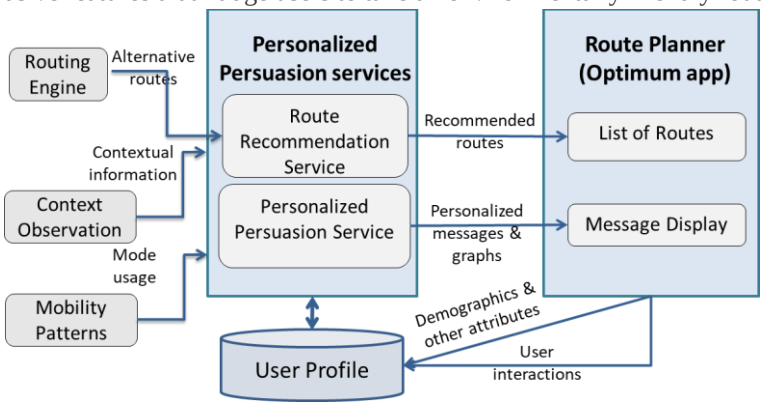


Figure 1: Overview of our personalized persuasion approach.

This paper focuses on the Personalized Persuasion Service (PPS), which implements the persuasive design elements described in Section 3. The aim is to persuade the user to take a particular route which is considered as optimal by the RRS. This route is the target for user persuasion. The PPS is called by the route recommendation service in order to attach a personalized persuasive feature to the route targeted for user persuasion. Persuasive features include tailored persuasive messages and graph-based visualizations.

The design and implementation of persuasive messages is described in our previous work [16]. The messages are presented in the route selection screen of the OPTIMUM mobile app as shown in part 1 of Figure 2 and implement the persuasive strategies of self-monitoring, comparison, suggestion and rewards. We have defined 125 persuasive messages with each one implementing a single strategy. Multiple messages have been designed per persuasive strategy, while all of them are context-aware, in the sense that they are valid for specific contexts. For example, the message "Today it's sunny! Take the opportunity to walk" will be presented to the user when the context "Nice weather" is active. The contextual elements used in our pool of messages capture the context in which the travel behaviour takes place. We have defined eight binary contextual variables (i.e. their value can be true or false) as follows: i) three variables based on personal travel behaviour characteristics (increased car / public transportation usage the previous period, caused emissions increasing compared to other users); ii) three variables based on trip-related characteristics (the destination is in a biking or walking distance, the duration of the route is similar to driving); iii) one variable based on weather status (nice or bad weather); iv) one variable based on rewards for sustainable route choices. Note that for the calculation of trip related context variables, users provide their preferences, including a maximum walking and bike distance preference, as well the ownership of a bicycle during registration. The graph-based visualizations that communicate statistical information about users' mobility habits, aim to nudge users towards sustainable choices by employing social comparison and self-monitoring persuasive strategies.

[1] http://www.optimumproject.eu

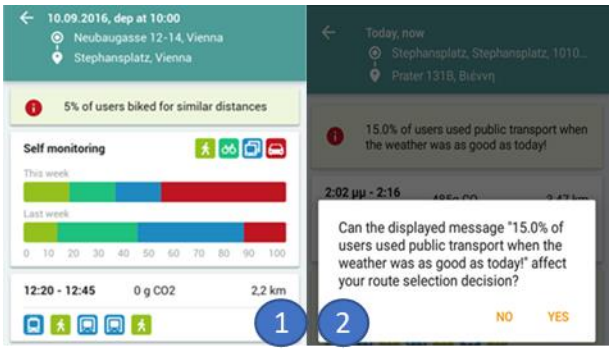


Figure 2: 1) persuasive messages and graph, 2) popup message to gather feedback on the impact of interventions.

The PPS tailors the persuasive messages and visualizations to the individual, with the aim to maximize the impact of the persuasive attempts. For personalizing the persuasive interventions, we rely on the concept of user persuadability, i.e. user susceptibility to different persuasive strategies. This is identified implicitly, i.e. without the need of explicit user involvement, and dynamically, i.e. after some attempts to persuade the user. The approach makes use of users' personality profiles as well as previous successful persuasive interactions of that particular user and other similar users. To address the so-called cold start problem according to which the system has not enough information for new users, we have developed a persuadability model that identifies users' susceptibility to different strategies on the basis of personality that is subsequently identified through a list of questions asked during user registration. For a detailed description please refer to [16].

Our approach combines persuasive visualizations with textual persuasive messages, in the sense that a user can be nudged to follow an environmentally friendly route by attaching to that route (i) a textual persuasive message, or (ii) a persuasive visualization, or (iii) both a textual message and a persuasive visualization. The persuasive strategies of self-monitoring and social comparison are implemented through both textual persuasive messages and persuasive visualizations, while the strategies of suggestion and reward are implemented solely through the textual persuasive messages. Only combinations of messages with persuasive visualizations that implement the same persuasive strategy are allowed, while the persuasive strategy to be used in each case is identified by considering the persuasive strategy that each user is more susceptible to.

The PPS controls the selection of the persuasive feature (i.e. messages or visualizations) that is presented to the user through the mobile application. Part 1 of Figure 2 depicts a sample screen where both a textual message and a persuasive self-monitoring visualization are presented with the aim to nudge the user to bike. As already mentioned, the user could be presented with similar screens with the textual message only or the persuasive visualization only. Moreover, our approach integrates mechanisms that adjust the intensity of persuasive interventions. This is performed with the use of a message throttling function that prevents displaying the same message or similar messages several times in a row, especially if they receive negative user feedback about their persuasive effect. A similar approach was followed for persuasive interventions employing visualizations. In addition, the intensity of interventions is adjusted according to the user trip purpose that is automatically inferred by the application. An indicative example is the following: suppose a user who mainly uses her/his car for everyday transportation. The intensity of persuasive interventions should be low for this user when traveling to work, whereas it could be higher for leisure trips as the chances of switching to more environmentally friendly modes should be higher in that case. For inferring the trip purpose of the user, we make use of information from external services which include Foursquare, and a location detection module which can automatically infer the users' home and work address based on previous visit patterns, in case such information has not be explicitly provided by users through the settings page of the route planning app.

5 Results

To test our approach and the effectiveness of the proposed persuasive design elements we setup a pilot study in three European cities which allowed us to investigate the mobility behavior, the travel mode choices as well as the attitudes and experiences of participants while using the OPTIMUM route planning app. During the pilot study, users from the city of Vienna, Birmingham and Ljubljana installed and used the OPTIMUM application to plan their everyday urban trips. The duration of the pilot study was 12 weeks, from May to July 2018. Altogether 72 participants took part in the study. They were equally distributed regarding gender (35 male, 35 females, 2 N/A), and the age range was between 17 and 66 years (mean 34.7, SD 11.9).

Users participated in two hours kick-off workshops where the topic of our study in general was presented. The workshops were performed in groups of up to nine people and with given dates. Users were chosen by their possibility to join a workshop, their gender, age, and car usage. During the kick-off workshop participants downloaded and installed the OPTIMUM app from Google Store on their mobile phone. Also, the subjects filled in an online questionnaire before the pilot. In this pre-pilot questionnaire, we asked participants to provide details on their travel preferences and past travel behaviour.

In addition, fifteen telephone interviews with selected users who actively used the OPTIMUM were conducted during and after the pilot to get a detailed status update. Interviews lasted approximately 30 minutes. Topics covered during the interviews included the general impression of the app, user experience related topics, opinions about persuasive messages and their influence on travel mode choices and personal environmental awareness. Finally, closing workshops were conducted after 12 weeks. During these workshops, participants filled in a post-pilot online questionnaire. The post-pilot questionnaire contained questions regarding usage and user experience of the app, the usefulness of the persuasive interventions as well as their influence on participants' actual travel behaviour and environmental awareness.

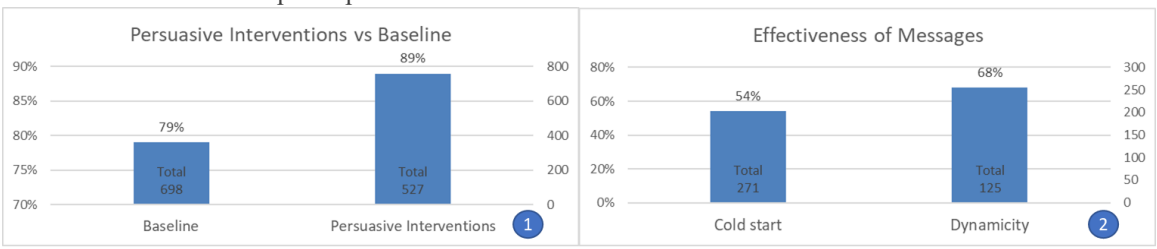


Figure 3: Percentage of route requests where users check the details screen of routes containing sustainable modes of transport (including walking, bicycle and public transportation) 1) in the baseline vs personalized persuasive interventions periods, 2) in the case of cold start vs dynamic approach.

Both quantitative and qualitative data were collected to gather as much insight as possible. For the data analysis, we took into account logged user interactions with the OPTIMUM app, the mobility and the personality type of participants and responses to the questionnaires. In order to better understand the behaviour and usage patterns of the users several means to collect data regarding these aspects were implemented. Data collection and analysis focused on the persuasive messages data and the GPS data. Specifically, whenever the user searched for a possible route with the app, the requests were collected and logged over the time of the pilot, thereby providing us the possibility to analyse users' activity about their route requests. In addition, the app automatically recorded the GPS-tracks and sensor information of the users' phones, and this data was then used to automatically impute the travel modality of the users.

In order to investigate the effect of "*DE1: Persuasive interventions specifically addressing each user and tailoring persuasive features to the individual*", we measured the impact of our personalized persuasive framework on behavioural change. Commonly this is measured by monitoring users' behaviour before and after the use of the persuasive technologies in order to observe whether users change their actions. However, as it is not always possible to measure or accurately determine actual behavioural change, we measured whether users viewed the details screen of routes containing sustainable modes of transport (including walking, bicycle and public transportation) in cases where users received and in cases where users did not receive persuasive interventions. In order to be able to measure this we had a baseline period of one week

when users did not receive persuasive interventions. Part 1 of Figure 3 shows the percentage of routes for which users checked the details screen of a sustainable route in case they received personalized persuasive interventions and the baseline.

To investigate the effect of "*DE2: Incorporation of a learning component that allows the system to adapt the persuasive interventions towards sustainable travel choices based on previous experience*", we compared the impact of persuasive interventions in the cold start case compared to the dynamic approach. As mentioned in Section 4, in the cold start case, the service selects the persuasive strategy that is more appropriate based on the static user persuadability model, while in the dynamic approach the selection of the persuasive strategy is performed by considering previous successful persuasive interactions of that particular user and other similar users. Part 2 of Figure 3 shows that the dynamic approach can be considered more effective as users viewed the details of routes containing sustainable modes of transport in 66% of the cases when adaptation of persuasive interventions was used compared to 54% when the learning component was not active.

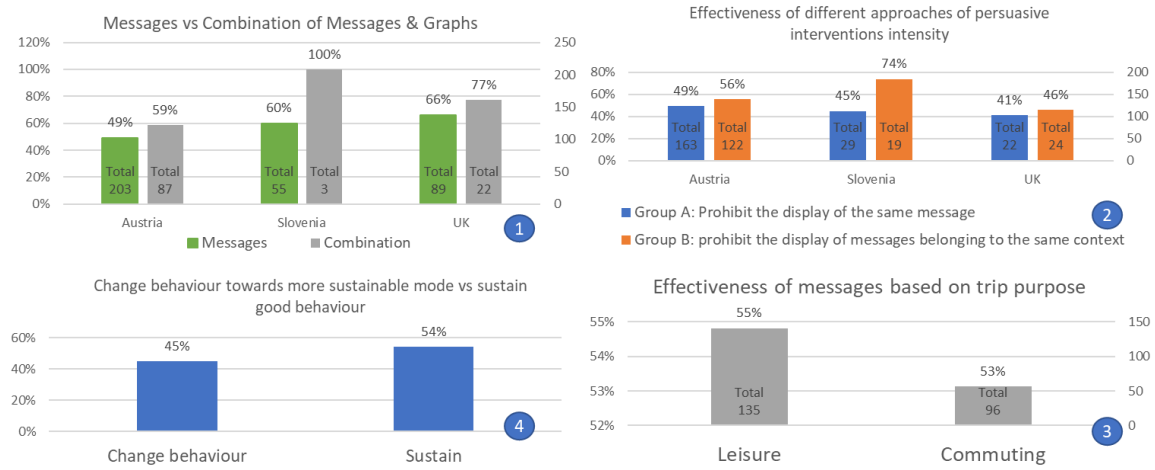


Figure 4: Overview of the effects of the design elements DE3 (1), DE4 (2), DE5 (3) and DE6 (4).

To understand the effect of "*DE3: Combining different means of communicating persuasive interventions (e.g. textual messages with graphs) to nudge users to make more sustainable travel choices*", we analysed the impact of persuasive messages versus a combination of messages and graphs. In this case, we measured the impact through a popup message which explicitly asked users whether the persuasive interventions affected their decisions (see part 2 of Figure 2). Part 1 of Figure 4 shows users' feedback for the two different approaches. The reported feedback shows that the combination of messages and graphs was more effective than messages alone only in all pilot cities. To study the effect of "*DE4: Adjusting the intensity of persuasive interventions*", we created two groups of users for which messages were throttled using two different methods. For the first group (Group A) we applied a throttling mechanism which prohibited the display of the same persuasive message before a number of persuasive attempts has taken place in the meantime. For the second group (Group B) we applied a mechanism which prohibited the display of messages belonging to the same context before a number of persuasive attempts has taken place in the meantime. Note that in our approach the unique persuasive messages (125) are significantly more than unique contexts (8) which results to a lower intensity of persuasive interventions for Group B users. Part 2 of Figure 4 summarizes the feedback from the users. We observe that the messages were more effective for Group B users where the intensity and repetition of the messages was lower. The reported feedback shows that the less repetitive the persuasive messages, the most effective seem to be.

To investigate the effect of "*DE5: Nudging users to sustain or intensify current green behaviour*", we compared the effectiveness of persuasive interventions depending on the target of persuasion in two cases. In the first case the system nudges users to follow more environmentally friendly routes that those they commonly take (e.g. a user who mainly takes his car will be nudged to follow public transportation or walk) whereas in the second case the system nudges users to continue following sustainable routes they already

take (e.g. a user who mainly takes public transportation will be nudged to continue taking public transportation). The results presented in part 3 of Figure 4 show that the effectiveness of persuasive interventions sustaining users' behaviour is 54%, while the effectiveness of persuasive interventions that nudge users to follow more environmentally friendly routes is 45%.

To identify the impact of "*DE6: Targeting persuasive interventions towards environmental friendly modes of transport mainly for leisure than commute trips*" we checked the effectiveness of persuasive interventions for different trip purposes. Part 4 of Figure 4 depicts the effectiveness of persuasive interventions for leisure and commuting purposes. The results show evidence that persuasive interventions are more effective for leisure purposes than for commuting.

6 Conclusions and Future Work

In this paper, we propose a set of persuasive design elements that can be incorporated into persuasive applications for urban mobility to nudge travelers towards more sustainable transport choices. The design elements have been taken into account for the design and development of an approach and related route planning application, which was used to empirically investigate how they affect the persuasiveness of interventions. Our results show evidence for a direct and measurable impact of the implemented persuasive interventions and the potential of persuasive features such as the persuasive messages and graphs to affect travel behaviour. Moreover, our results indicate that the personalization of persuasion can nudge users towards the selection of environmentally friendly routes.

We find evidence that the combination of messages and graphs was more effective than messages only which means that persuasive technology designers should combine persuasive messages with graphs to ensure greater persuasive impact. We also observed that the less repetitive the persuasive messages, the most effective they are. Moreover, we found that the effectiveness of persuasive interventions is higher for leisure purposes than for commuting. Based on these results, persuasive technology designers should provide solutions that are not too pressing and should consider the trip purpose as the interventions' effectiveness greatly depends on this factor. Furthermore, our results indicate that nudging users to select sustainable modes that they have already used is more effective. This means that a persuasive target should not be selected on the basis that a route is greener than the typical transportation choices of the user but also consider modes the user is familiar with and support him/her to sustain or intensify current green behaviours as well as to resist the temptation of using more unsustainable modes of transport. Since a behavioural change is rarely easy and often requires a gradual progression of small steps toward a larger goal, the persuasive target among two routes that are both more sustainable than the route which the user typically takes, should be the one that is more similar to the typical user choice.

However, we must note that the short time frame of the evaluation did not allow us to gather enough data which would allow to test the statistical significance of the proposed interventions. Future work should focus on delivering largescale and longitudinal evaluations in order to establish a solid opinion on the long-term impact of persuasive technologies at promoting desirable behaviours in the area of urban mobility in smart city contexts.

ACKNOWLEDGMENTS

Research reported in this paper has received funding by the EC project OPTIMUM (GA no. 636160-2).

REFERENCES

[1] Zlotnik H. (2017). World urbanization: trends and prospects. New Forms of Urbanization. Routledge, 43-64.

[2] Pucher J. and Dijkstra L. (2003). Promoting Safe Walking and Cycling to Improve Public Health: Lessons From The Netherlands and Germany. American Journal of Public Health, 93(9), 1509-1516.

[3] Birch E. (2014). A Review of Climate Change 2014: Impacts, Adaptation, and Vulnerability, Journal of the American Planning Association, 80(2), 184-185.

[4] Fogg B. (2003). Persuasive technology: Using Computers to Change What We Think and Do. Morgan Kaufmann, San Francisco.

[5] Wiafe I. and Nakata K. (2012). Bibliographic Analysis of Persuasive Systems: Techniques; Methods and Domains of Application. In 7th International Conference on Persuasive Technology, Linköping, Sweden, Adjunct Proceedings, 068, 61-64.
[6] Kimura H. and Nakajima T. (2011). Designing Persuasive Applications to Motivate Sustainable Behavior in Collectivist Cultures. PsychNology Journal, 9(1).
[7] Froehlich J., Dillahunt T., Klasnja P., Mankoff J., Consolvo S., Harrison B. and Landay J. A. (2009). UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems 1043-1052.
[8] Bothos E., Mentzas G., Prost S., Schrammel J. and Röderer K. (2014). Watch your Emissions: Persuasive Strategies and Choice Architecture for Sustainable Decisions in Urban Mobility. PsychNology 12(3).
[9] Holleis P., Luther M., Broll G., Cao H., Koolwaaij J., Peddemors A., Ebben P., Wibbels M., Jacobs K., Raaphorst S. (2012). TRIPZOOM: A system to motivate sustainable urban mobility. In 1st Int. Conf. on Smart Systems, Devices and Technologies, Stuttgart, Germany.
[11] Wells S. (2014). SUPERHUB: Towards argumentatively rich persuasive communication in large-scale, applied behaviour change systems. COMMA Demo Track, Scotland.
[12] Anagnostopoulou E., Bothos E., Magoutas B., Schrammel J. and Mentzas G. (2018). Persuasive technologies for sustainable mobility: State of the art and emerging trends. Sustainability, 10(7), 2128.
[13] Klecha L. and Gianni F. (2017). Designing for sustainable urban mobility behaviour: a systematic review of the literature. Smart Learning Ecosystems and Regional Development, Springer, Cham, 137-149.
[14] Sunio V. and Schmöcker J. D. (2017). Can we promote sustainable travel behavior through mobile apps? Evaluation and review of evidence. International Journal of Sustainable Transportation, 11(8), 553-566.
[15] Gabrielli S., Maimone R., Forbes P., Masthoff J., Wells S., Primerano L. and Pompa M. (2013). Designing motivational features for sustainable urban mobility. In ACM CHI'13 1461-1466.
[16] Anagnostopoulou E., Urbančič J., Bothos E., Magoutas B., Bradesko L., Schrammel J. and Mentzas G. (2018). From mobility patterns to behavioural change: leveraging travel behaviour and personality profiles to nudge for sustainable transportation. Journal of Intelligent Information Systems, 1-22.
[17] Halko S. and Kientz J. A. (2010). Personality and Persuasive Technology: An Exploratory Study on Health-Promoting Mobile Applications. In PERSUASIVE, 6137, 150–161.
[18] Heer J., Bostock M. and Ogievetsky V. (2010). A tour through the visualization zoo. Queue, 8(5), 20.
[19] Pandey A. V., Manivannan A., Nov O., Satterthwaite M., and Bertini E. (2014). The persuasive power of data visualization. IEEE transactions on visualization and computer graphics, 20(12), 2211-2220.
[20] Mylonopoulou V. (2016). Using Social Visualisation as a Tool in Persuasive Technology to Cause Behavioral Change. Available at http://www.chessitn.eu/2016/03/14/using-social-visualisation-as-a-tool-in-persuasive-technology-to-cause-behavioral-change/
[21] Allen M., Bruflat R., Fucilla R., Krame M., McKellip S., Ryan D. J. and Spiegelhoff M. (2000). Testing the persuasiveness of evidence: Combining narrative and statistical forms. Communication Research Reports, 17(4), 331-336.
[22] Foster D., Lawson S., Wardman J., Blythe M. and Linehan C. (2012). Watts in it for me?: design implications for implementing effective energy interventions in organisations. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, 2357-2366.
[23] Petersen J. E., Frantz C. and Shammin R. (2014). Using sociotechnical feedback to engage, educate, motivate and empower environmental thought and action. Solutions, 5(1), 79-87.
[24] Mangiapanello K. A. and Hemmes N. S. (2015). An analysis of feedback from a behavior analytic perspective. The Behavior Analyst, 38(1), 51-75.
[25] Berkovsky S., Freyne J. and Oinas-Kukkonen H. (2012). Influencing individually: fusing personalization and persuasion. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(2), 9.
[26] Gunn H. F. and Pattarathep S. (2007). An Introduction to the valuation of travel time-savings and losses. Handbook of Transport Modelling: 2nd Edition. Emerald Group Publishing Limited, 503-517.
[27] Shires J. D. and Gerard C. (2009). An international meta-analysis of values of travel time savings. Evaluation and program planning 32(4), 315-325.

Mobility as a Service

A valuecentric approach to design

Rob Christiaanse*
TUDelft
Delft, South-Holland
EFCO BV
Amsterdam, North-Holland
r.christiaanse@efco-solutions.nl

ABSTRACT

MAAS as a mobility model leans on the idea that a gap between private and public transport systems needs to be bridged as well as on a city, intercity, national and supranational level. The current situation is felt problematic due to the fragmented tools and services often organized in silos to meet a traveler needs to undertake a trip. One of the major concerns designing any platform system like Mobility as a Service is where to start modeling and how to express the notion of the platform system in some language that is understandable for all stakeholders of the platform system. Understandability buttresses the expectation of stakeholders whether some design will probably implement the intended platform services enabling users to actually buy and or use the platform system for what ever purpose. Building on the economic theories of two-sided markets and mechanism design we introduce the concept of value nets extending the Contract Protocol Net. Value net modeling offers a precise abstract representation which provides in the detailed informational requirements in a canonical form and it connects i.e. implements the abstract notion of Service Oriented Architecture characterizing systems without loss of crucial informational elements.

CCS CONCEPTS

• **Applied computing → Enterprise computing**; **Enterprise computing infrastructures**;

KEYWORDS

Mobility as a Service, two-sided markets, mechanism design, value net modelling, smart contracts

ACM Reference Format:
Rob Christiaanse. 2019. Mobility as a Service: A valuecentric approach to design. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317050

*Corresponding author

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317050

1 INTRODUCTION

Making city governance more efficient, citizens happier, business more prosperous, and the environment more sustainable through the deployment of systematic integration of technical infrastructures that rely heavily on advanced data-processing is coined as smart city [20]. Smart mobility is one of the smart city dimensions that includes dynamic traffic- and congestion-management, cooperative intelligent transport services (C-ITS), connected autonomous driving (C-CAD) and mobility as a service (MAAS) [5] [6]. Essentially MAAS makes a multi modal lifestyle feasible from the perspective of the citizen introducing a novel transport paradigm in which it is no longer needed to own a car by offering travelers access to a range of transport alternatives packaged to personal preferences. These packages combine offerings from different transport services like public transport with shared bikes, or public transport services with on demand responsive transport, car sharing systems, car sharing in combination with sharing bikes, or car rental on a day-to-day basis or vehicle rental paying for the actual use and so on. It needs no elaboration that the notion of sharing viewed as a (business concept) is one of the key notions that underlie or even underpin smart mobility as a dimension of smart city notions and concepts. Within a governmental policy context the main question or goal is can MAAS lead to behavioral changes able to bring about (1) a better product for the citizen, (2) Social inclusion, (3) Congestion reduction, (4) Accessibility and last (5) Sustainability [9]. Indeed when we come up with a design than we have to address these questions. The key issue underlying a design that implements MAAS service providers enabling citizens to actually buy the MAAS service is where to start modeling and how to express the notions in some language that is understandable for all stakeholders of the MAAS service system. MAAS as a mobility model leans on the idea that a gap between private and public transport systems needs to be bridged as well as on a city, intercity, national and supranational level. The current situation is felt problematic due to the fragmented tools and services often organized in silos to meet a traveler needs to undertake a trip. In this paper we introduce the concept of value to model MAAS services [1]. Through the value model coined as **value net modeling** we have a mechanism that ensures that design decisions are made consistently safeguarding that the underlying social and technical notions are aligned, consistency is warranted and a privacy preserving mechanism is ensured by means of the design of a two-sided platform like MAAS.

[1] In our exposition we lean heavily on the concepts explained in [17] [16] [14] [18]

2 RESEARCH APPROACH

In a scientific context testing a hypothesis means confronting statements about an assumed relationship between phenomena with empirical facts. In a design context the terms testing and hypothesis tends to shift in meaning, a design assumption is not a matter of being true or false but given a particular context it is a matter of being the best solution based on vision and believes [19]. As a consequence very different logics of discovery may be at work in design practices, and the way they are mixed may vary form case to case, from situation to situation, from context to context and so on. Whatever mix or configuration of elements, we will always need (good) theories to account for what happened. Theories are constitutive for every design just because we need to understand why some things do work and other things do not work or will never work. We must explain in advance why. Put in other words, we need a explicit interpretation of what is constituted as the tacit understanding, just displayed i.e. showed in practice [23]. Some authors state that "testing a design hypothesis is inextricably bound up with the (ethical) normative framework of society and with its epistemological principles" [22].

2.1 Empirical grounding

This research is initialized and inspired by an European tender convoked by the dutch Ministry of Infrastructure and Watermanagement [9] in 2017. A consortium of two professional firms in the automotive and smart card technology sector developed a commercial proposition with respect to MAAS services. This research has contributed to the tentative design idea buttressing the business model of the MAAS service proposition as one of the desired outputs of design science research process model in [38].

2.2 Purpose and process

Following Lewis [3, 31] good representations of meaning are only possible when at the same time a statement is made on how the representation of the meaning is used in for example communication and inference. Structure defined as an assembly of components should always be studied in tandem with an associated process, whatever this process may be. The same duality seems valid for the *notion* of information or enterprise systems design. What does information do for each process, and can we find one abstract level of representation that stays away from the details of implementation [3]? This research is in the realm of Design Science Research [30] and is to be characterized as Design Theory. Our aim is how to come up with a design of MAAS services without fully answering why the prescribed model as actions should work. In this respect this research coined as design relevant explanatory/predictive theory (DREPT) augments the "How" part or question with explanatory information on "Why" one should trust the proposed design actually will work.The key point is that the explanatory information is obtained using kernel theories. Kernel theories are established theories form social sciences, economics, mathematics, computer science, logic and so on. We are interested in theory building on how to design effective and efficient MAAS services, of which this may be interpreted as experimental scientific investigation. In this paper we develop as a first step a methodological construction of the MAAS artifact as the object.

3 MAAS

3.1 The notion of Mobility-as-a-Service

There are many ideas, concepts and definitions for describing MAAS services. Intelligent Transport Systems Finland (ITS Finland) view Mobility as a Service (MAAS) as a mobility distribution model implemented as a single platform by a single service provider that orchestrates each individual transport service component to meet "customer's end-to-end service expectations" [27]. The dutch Ministry of Infrastructure and Watermanagement define MAAS as "the provision of multi-modal, demand-driven mobility services, whereby customers are offered tailor-made travel options via a digital platform (e.g. Mobile Apps) based on real-time information, including payment and the handling of transactions [9]. Kamergianni et al define MAAS as a user-centric, intelligent mobility distribution model in which all mobility service providers' offerings are aggregated by a sole mobility provider, the MAAS provider, and supplied to users through a single digital platform [28]. Intuitively these definitions are precise enough for a rough understanding of the notion. It is less clear which concepts buttress MAAS precisely, but there is one aspect that concerns us most. In the definitions aforementioned it is expected or assumed that MAAS is implemented by some sort of single platform by a single service provider. We think this monolithic view is too limited and neglects the distributive nature of the notion of MAAS and therefore limits design and solution space as we will address later on in this paper. MAAS as a mobility model leans on the idea that a gap between private and public transport systems needs to be bridged as well as on a city, intercity, national and supranational level. The current situation is felt problematic due to the fragmented tools and services often organized in silos to meet a traveler needs to undertake a trip. So to speak MAAS can be coined as a bridge defined as a type of computer network device i.e. a mechanism that provides interconnection with other bridge networks that use the same protocol. More specifically technically a bridge network works at the data link layer of the Open System Interconnect (OSI) model, connecting two different networks together and providing communication between them [42]. Conceptually the role of data evolved from a syntactical representation of an object to an economic valuable representation of an object. In the situation data can be decoupled from specific hardware and software implementations, then data turns into an independent economic good. This shift opens up new opportunities and facilitates i.e. enables digital business models like MAAS [8]. Here we see the novelty of MAAS in creating value for end users like customers, travelers, or other participants in the smart-mobility system. The novelty consists of linking networks to create network effects by performing a bridge function between two or more separate networks. In summary two aspects are paramount to understand MAAS namely the decoupling of data from specific hardware and software implementations and the coupling of networks

3.2 Characterizing Mobility-as-a-Service

It needs no elaboration to see that when data also become economic goods that MAAS services are both technical as economical and sociological. One way to describe MAAS as a mechanism bridging

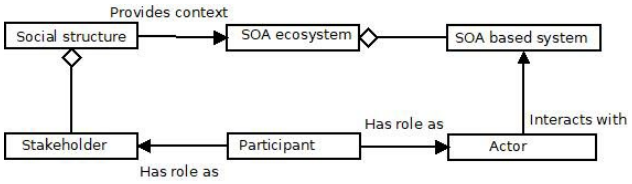


Figure 1: ECO systems model

networks is to use service oriented architecture (SOA) as an architectural paradigm to describe for decision purposes the abstract architectural elements in the domain of interest, here smart-mobility, independent of the technologies, protocols and products to use to implement a specific solution for the domain [34]. MAAS services as a system can not be fully understood by simply decomposing the system in its parts and subsystems. This is certainly the case when many autonomous parts of the system are governing the interactions among them. By understanding the situational context in which MAAS, as a bridge, functions and the participants using the MAAS services involved in making the MAAS bridge function work, then we can come up with a design and proposed governing design principles of which we are confident that the proposed design of MAAS services will actually have the potential to work in an appropriate manner. Hence this view on MAAS can be conceptualized as an ecosystem viewed as a whole defined as a network of discrete processes interacting with communities (one of more peoples), creating, using and governing specific services. In short the SOA is a paradigm for(1) exchange of value between independently action participants, (2) participants have a legitimate claim to ownership of resources that are made available within the MAAS ecosystem and (3) the behavior and performance of the participants are subject to rules or engagement which are captured in policies and contracts [14, 34]. Conceptually a SOA ecosystems model is depicted in fig1

From a smart-mobility point of view we are interested in the right part of the model where actors interact with SOA based systems with some goal in mind. MAAS services implement the means to attain the actors goal, which is to be able to travel from A to B.

3.3 One sided versus two sided markets

The mechanism bridging networks is often addressed as platform. We will use the term platform exclusively to address two-sided markets opposed to traditional one-sided markets. The notion of two-sideness relate primarily to the theories of network externalities that borrows the assumption that there are non-internalized externalities among end-users. More formally the market for interactions between two sides (i.e. the buyer en the seller) is one-sided if the volume **V** of transactions realized on the platform depends only on the aggregate price level **a**, where a is the total price of the price a buyer pays a^B and the price the seller pays a^S using the platform. The market is said to be two-sided when volume **V** varies with a^B while keeping **a** constant. The utility of a buyer U^i is equivalent to the difference of the average benefit b^i per transaction and the usage fee a^i per transaction times the number of members, i.e. the sellers, on the other side plus the fixed benefit for the buyer B^i minus the fixed benefit for the seller A^i [35]. Now it is easy to see that coupling networks generate value by internalizing the externalities by means of a pricing mechanism i.e. a value function complementing the pricing mechanism buttressing the service exchange mechanism itself. Indeed a platform reduces uncertainties relating to transaction costs [37] and due to the volume effect one can expect that service prices themselves will drop. For example sharing taxi's will reduce the price per trip per customer. The user-customer can have a plethora of motivations to share a cab, take a bicycle, use public transport systems or a combination of two or more modalities. One of them is simple we feel that we have to contribute in lowering our CO2 footprint. Indeed MAAS is usercentric and demand driven. These behavioral aspects need to taken into account designing smart-mobility systems like MAAS as an SOA ecosystem.

3.4 MAAS mechanistically viewed

The platform bridging networks viewed as a mechanism is a mathematical structure that models institutions through which for example economic activity is guided and coordinated. When we design a platform like MAAS than we need to know what the problem of mechanism design is [26]: Given a class Θ of environments, an outcome space Z, and a goal function F, find a privacy preserving (i.e. a decentralized) mechanism $\pi = (M, \mu, h)$ that realizes F on Θ, where M is the message space, μ denotes the (group) equilibrium message correspondence $\mu : \Theta \mapsto M$ and h denotes the outcome function $h : M \mapsto Z$. The key insight of [25] was that information about the environment, facts that enable or constrain possibilities are **distributed** among agents. But pay attention when an agent is not able to observe some aspect of the prevailing environment, than the agent does **not** have the information to guide his or her actions, unless the agent is communicated to by another agent who was able to observe. Notably an agent is not able to observe the private information of another agent so **trust** must be accounted for. By means of a verification scenario which separates the process of finding an equilibrium from recognizing an equilibrium it is possible to design incentive compatible mechanisms which occurs when the incentives that motivate the actions of individual participants are consistent with following the rules established by the group or set by the institution like the MAAS provider. Simply put, in a verification scenario each agent reads the announced message by saying yes or no. The proposed outcome is judged acceptable if and only if the agents' responses are affirmative. The message exchange process consists of three elements first a message space M, second a group equilibrium message correspondence μ, denoted $\mu : \Theta \mapsto M$ and third outcome function h, denoted as $h : M \mapsto Z$. A message space M consists of the messages available for communication. Messages may include formal written communication like contracts among buyers and sellers, accounting reports, production statistics, emails and so on.

3.5 Contracts and value-nets

From a behavioral point of view a contract can be defined as a statement of intent to regulate behaviors among agents [11]. Thus a contract provides in a comprehensive set of both constitutive and regulative rules (i.e. norms) to construct a social reality[12].

Constitutive rules introduce abstract classifications of facts and entities or they describe the (legal) consequences of actions on the normative system, where regulative norms provide notions of conditional obligations with an associated sanction (or reward). In essence a contract depicts the content, structure, and governance of transactions [4, 43]. The transaction content refers to the goods, services or information that is to be exchanged and the resources and capabilities that are required to enable the exchange of goods, service or information delivery. The transaction structure refers to the collaborating organizations participating in the exchange and describes how collaborating organizations are interlinked, inclusive the sequencing in which delivery will take place en the adopted mechanism for enabling transactions. Transaction governance depict the way in which information flows, resources, goods and services are controlled by participating organizations in the exchange as well as other stakeholders like regulators and societal agencies. Transaction governance embodies the legal and other regulatory aspects of organization as well as the incentive structure for the participants in the transactions engaged.

Contracts are a legal,sociological and economical devises to align goals, intentions, desires and obligations among participating organizations engaged in a exchange of goods, services, information and money. If we accept the behavioral view of a contract, than we need to address all three transaction aspects of a contract simultaneously. So if we change the transaction content, we have to address the transaction structure and the transaction governance, if we change the transaction structure we need to address the transaction content and transaction governance, if we change the transaction governance we have to address the transaction structure and the transaction content. Needless to say that balancing the three aspects is a complex endeavor. Consequently a contract is not just a set of clauses earlier for convenient purposes defined as a set of clauses C = {C1, . . . , Ck}, but contains specific interrelated sets of clauses referring to content, structure and governance buttressing the transaction agreed upon between agents as users, organizations and institutions. As such a contract serves as a legitimate institution in organizations of agents [11, 12, 18]. Normative systems, like a design of MAAS, prescribe how the agents ideally should and should not behave. Norms in general provide in the possibility that actual behavior may at times deviate from the ideal situation. Given the notion of constitutional and regulative rules, regulative rules i.e. norms can be mapped onto the transaction structure, whereas constitutive rules covers the transaction content and the transaction governance. The transaction structure provide in the notion of conditional obligations (or permissions and prohibitions) with an associated sanction (or reward). The associated sanctions (or rewards) refer to the pre-conditions depicted in the transaction content and or to the post-conditions depicted in the transaction governance. Economically and thus **computationally** the price mechanism actually serves as a balancing mechanism across the transaction content and the transaction governance on one side and the transaction structure on the other side. Stated otherwise, if the transaction content and the transaction governance are equal to the transaction structure than the outcome of the price mechanism, that is the agreed upon market price, actually reflects the mutually agreed upon conditions between participating organizations reflected by the actual public communication flow. Stated otherwise,

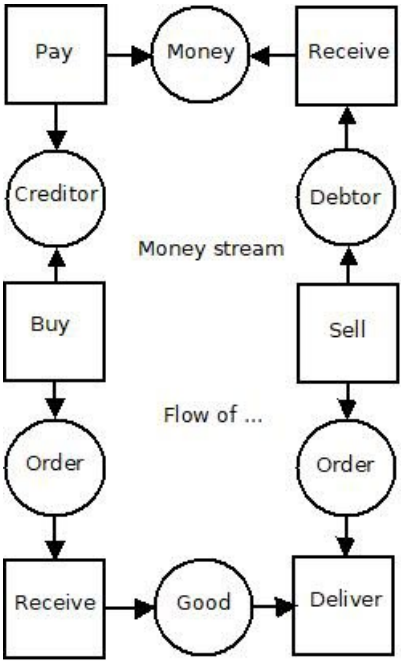


Figure 2: Canonical form of a Valuenet

if we use the mutually agreed upon price between participating organizations than it must be the case that the transaction content provides in the necessary specifications of the pre-conditions, and that the transaction governance specifies the necessary post-conditions in the case the conditional obligations (and permissions and prohibitions) and or preconditions are not met in such a way that the specifications are actually equal to the drawn upon transaction structure described as conditional obligations (and permissions and prohibitions) upon which the participating organization must comply with.

We stated that from a economically and computational point of view that the price mechanism it self or the outcome of the applied price mechanism actually serves as a balancing mechanism across the transaction content and the transaction governance on one side and the transaction structure on the other side. This stance needs further clarification. Transactions coined as economic events effect the financial position of an enterprise and stem per definition from contractual arrangements between agents. Agents in the marketplace agree upon *the conditions the goods and or services to be delivered* and *the market price to be paid in return*. Economic transactions as described above can be formally modeled as a value cycle which is interlinked to the business model of the enterprise i.e. the revenue model [15, 16]. When used in computer science, often the purpose of these models is to analyze the representations of actions and events in a business process, and study their well-formedness. Consequently an intra-organizational communication flow i.e. business processes can be modeled as a value net which provides a top-level view of an enterprise that focuses on the economic events equivalent to the value cycle of an enterprise.

3.6 Value-nets: a value centric approach to two sided platforms

More formally the canonical form of a value net is modeled with a dimensioned Petri Net, extended with a place sign and a valuation, with some special structural characteristics that make it a good representation of the intra-organizational process structure [15, 41]. For completeness purposes we give the following definition.

Definition 3.1. A *Value Net* is a tuple (P, T, F, B, s, v) with

- P a finite collection of places,
- T a finite collection of transitions,
- $F, B : P \times T \to \mathbb{N}$ the net's incidence matrices
- $s : P \to \{\text{asset}, \text{liability}\}$ the indicator for each place's sign, and
- $v : P \to \mathbb{R}$ a valuation

Additionally a value net has the following structural properties:

- The net is cyclic in the following sense. There is a special place labeled *money* from which it is possible to reach every other node in the model and that can be reached from every node in the model. Put differently, every node in the model is on a path from *money* to *money*.
- The net is divided in a money part and a flow of part. Place *money* is in the money part. The transitions in which the input places are in the money part and the output places in the flow of part are called *buy* transitions. The transitions in which the input places are in the money part and the output places in the flow of part are called *sell* transitions.

We use subscripts to differentiate the money and the flow of part. So, partitioning $P = P_{mon} \cup P_{flowof}$ splits the places into the money part and the flow of part. Transitions in event money are between places the money part and transitions in event flow of between places in the flow of part. A event buy transition consumes value from the money part and produces value in the flow of part. An event sell transition consumes value from the flow of part and produces value in the money part.

As we have seen in 3.3 a distinction is made in one sided and two sided markets. To model two sided markets we have to model one-sided markets first. To model a one-sided market we extend the value net as depicted in figure 2 with a duplicate. The result is depicted in figure 3 coined as the Double.

On the left-hand side we have the original value net representation of the intra-organizational process structure of an organization. On the right-hand side we see the duplicate of the left-hand side representing the intra-organizational process structure of the buying organization. This makes the left-hand side the selling organization or agent and the right-hand side the buying organization. Markets separate sellers and buyers. In the case the seller and buyer agree upon the contract; that is seller and buyer agree about the price of the goods, service or information, agree of the quality of the goods, service or information and agree about the conditions like user rights, warranty en ownership and so on; than money is transferred and goods are delivered. The former depicted in the top arrow from money to money and the latter depicted as a bottom arrow from good to good transferring owner, user rights and obligations.

Now we can model the two sided market by adding a value net model similar to the value nets of the seller and buying organizations except we have to model the notion of a platform coupling networks as defined in 3.2 and 3.3. The key characteristic of platforms are **search** and **find** capabilities. The result is depicted in figure 4.

As we see in figure 4 coined a the Triple the right-hand side buys search and find capabilities. Mobile apps for example instantiate these capabilities to accomplish specific tasks. Notice that two top level arcs are added representing the money flow from the left-hand side the seller is willing to pay the price a^S using the platform and the money flow from the right-hand side the buyer is willing to pay the price a^B Remember that for interactions between two sides (i.e. the buyer en the seller) the market is said to be two-sided when volume **V** varies with a^B while keeping **a** constant. At the bottom side of figure 4 two dashed arcs are added representing the found request of the buyer. This result is precisely the model treated in section 3.4. specifically the notion of **trust** and the necessary **verification scenario**. The key question is who governs the verification scenario and how to do we guarantee or secure the agents' privacy.

3.7 Trust in platforms or privacy preserving mechanisms in MAAS

The Contract Net Protocol (CNP) is a task allocation protocol that facilitates negotiation between bidders and an auctioneer in a Multi Agent System to form a contract. Bidders for example, here the buy-side want to interact with one or more sellers, to update their bids until their bids are accepted. In this respect it is important that the Contract Net Protocol terminates correctly [10]. Time sequence diagrams are often used for depicting the communication flow a like activity diagrams and so on. The CNP covers the communication between buy and sell and counts as an instantiation that works at the data link layer of the Open System Interconnect (OSI) model, connecting two different networks together and providing communication between them [42]. From a design point of view is is easy to extend the Double with communication as depicted in figure 5.

It is easy to see that the CNP facilitates the process of bidding or dealing with quotations which buyer and seller eventually agree upon. From a control point of view CNP like protocols are limited because the actual payment and delivery are not part of the the CNP model. This is why the value net modeled in the canonical form with a dimensioned Petri Net gives us the correct insight. After agreement of the contract the credit department of an organization needs to know what has to be paid and what has been bought at what conditions. On the other side the organization needs to know what is to be received in what quality. The same reasoning is also applicable for the seller. A verification scenario fulfills these type of checks and balances. The interfaces (API) to creditor, debtor and order need to be accounted for otherwise the verification scenario will become unduly costly [17]. Potentially it will deteriorate the data quality of the system and addresses concerns about the completeness and the accuracy of the data in the accounting, control and audit domain. It needs no elaboration that trust in platforms are interlinked with the data quality and the representational faithfulness of the data registered with a platform as defined in 3.1.

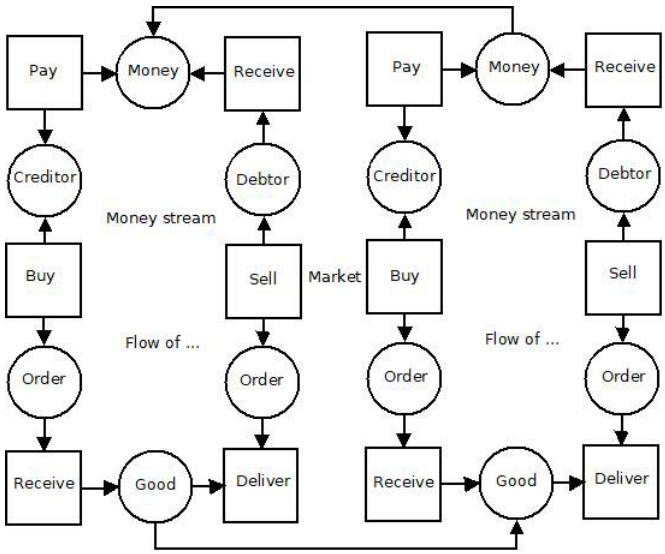


Figure 3: Double

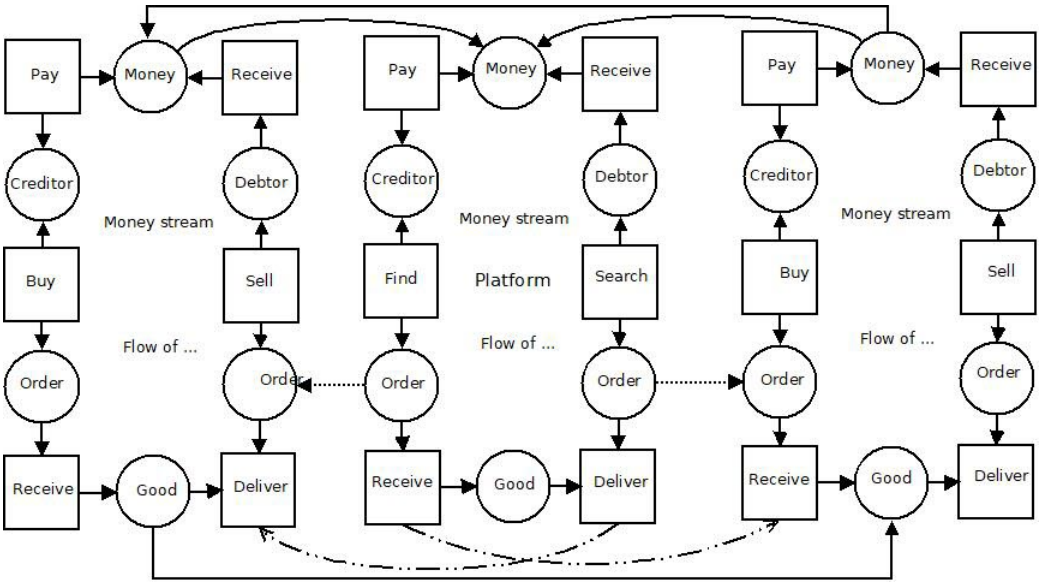


Figure 4: Triple

[13] Closely related to the data quality concerns in terms of completeness and accuracy is the notion of privacy of the agents using information systems like platforms as MAAS. From a regulatory point of view one needs to take into account the GDPR on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). In this paper we assume that the non-functional requirements will be met. There is one aspect which is very relevant for maintaining the privacy of users and this concerns the necessary identification, authentication and authorization procedures which buttress all security standards. Additional to the verification scenario the verifier needs to validate whether a user is authorized i.e. has the permission to act upon the data he or she has access to [24]. As we have addressed in section 3.5 it must be the case that the transaction content provides in the necessary specifications of the pre-conditions, and that the transaction governance specifies the necessary post-conditions. This is why attribute-based policies needs to be integrated with Role-based Access Control (RBAC) [21, 39]. In section 3.1. we addressed our concern to the expectation or the assumption that MAAS is implemented by some sort of single platform by a single service provider. We stated that this monolithic view is too limited and neglects the distributive nature of the notion of MAAS and therefore limits design and solution space. We need to clarify this. As we have seen

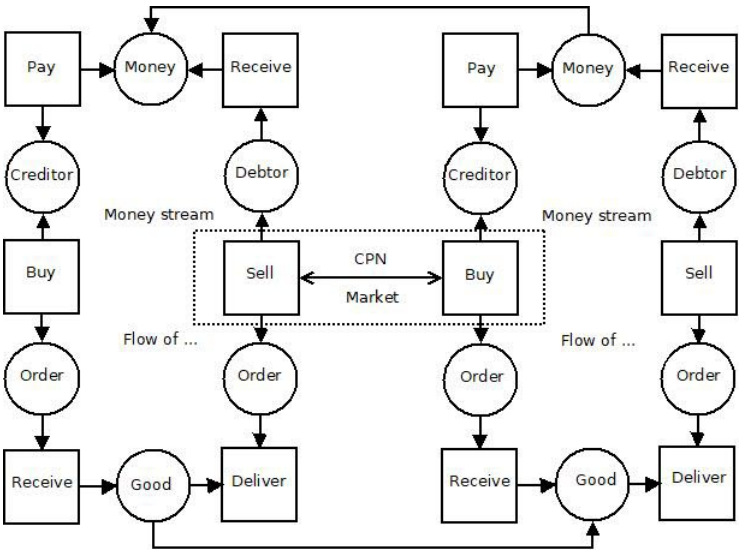


Figure 5: DialogDouble

in section 3.4. and 3.5. a (smart-mobility) service platform is a fully decentralized in nature so the design question is whether a central authority is warranted. Models with a strict central authority stimulate monopolistic or oligopoly behavior as studied extensively and demonstrated by market failure theories and social choice theories [7, 40]. It cannot be debated that trust and justice needs to be established and maintained. In communication networks receivers of information want to ascertain its origin when needed, they want to verify whether a message or some other information has not been modified during transit and they want to be sure that senders of the message are not able to deny that the message was send at all [36]. These are vital requirements known as authenticity, integrity and non-repudiation. Recently crypto currencies like bitcoin are implemented with block-chain technology are on the rise because of the combination of being a currency and a payment system as well it has some attractive features enabling local privacy preserving properties. Nakamoto observes that no mechanism exists to make payments over a communications channel without a trusted party [32]. Nakamoto states that an electronic coin can be defined as a chain of digital signatures. Each owner of the coin transfers the coin to the next owner (to be) by digitally signing a hash of the previous transaction (from the viewpoint of the owner to be) with his private key (current owner) and the public key of the next owner (the owner to be) and adding these (the hash plus signature and keys) to the end of the coin. A payee can verify the signatures to verify the chain of ownership. Because of the **public nature** one problem had to be resolved in the case users of the crypto currency payment scheme is that users face the risk of double spending. A user (the payee, that is the owner to be) needs a way to know that the previous owners (the current owner and their successors) did not sign any earlier transactions. So the design of the payment scheme has to provide in a mechanism that (positively) confirms the absence of âĂIJdouble spend coinsâĂİ in the former transaction (chain). The mechanism must be aware of all transactions and provides in the need due to the absence of a trusted third party like banking institutions. The key idea buttressing the bitcoin is based the notion of a time-stamp server, proof-of-work and an incentive mechanism. The transactions to be confirmed are placed in a hashed time stamped block which includes the previous time-stamp in its hash, forming a chain coined as the blockchain. Notice that the implementation of crypto currencies covers the actual payment of the services, as the CNP protocol only covers the bidding over services for the best price.

Combining digital signatures with public-key cryptography we get a protocol that combines the security of encryption with the authenticity of digital signatures. The protocol is as follows, where **A** and B denotes the agents, **S** denotes signing, **Pka** denotes the private key of A, **Pkb** denotes the private key of B, **D** denotes decrypt, **Pua** denotes the public key of A, **Pub** denotes the public key of B and **M** is the message containing the conditions of the contract.

- A signs a message with his or her private key: A-SPka(M)
- A encrypts the signed message with Bs public key and sends it to B : A-EPukb(A-SPka(M))
- B decrypts the message with his private key: B-DPkb(A-EPukb(A-SPka(M)))= A-SPka(M)
- B verifies with As public key and recovers the message : B-VPuka (A-SPka(M))=M

Notice that the contents of M itself can be encrypted to some standard. The main design question is what should be the contents of the message and for what purpose is the data exchanged and registered in information systems. Exchanges are by definition **reciprocal** in nature and come in a large variety of what we coin as means like signed contracts, shaking hands et cetera. For example signing a contract by both parties is performative in nature; by the act of signing, we communicate that the exchange is done. Hence a signed contract affords exchanging. An affordance establishes the relationship between an object or an environment and an organism

here a (human) agent through a stimulus to perform an action. In our example the stimulus is the signed contract and the detectable change in the external environment. It needs no elaboration that our protocol as given above needs to be extended by a second massage to make the protocol reciprocal. Indeed the given protocol is one-sided i.e. unilateral by nature. We need a bilateral contract in which the parties exchange mutual promises. Bilateral contracts are commonly used in business transactions; a sale of goods is a type of bilateral contract. The seller promises to deliver the goods or services where the buyer promises to pay for the goods by providing the seller with the indebted amount of anything parties have agreed upon. Hence we need two signatures one of the seller and one of the buyer. Logically the information flow of a bilateral contract can be depicted as a directed graph partitioned in two separate flows. Mathematically a value net is a bipartite graph so the bilateral contract modeled as information flows i.e. the canonical model for the communication protocol has the same properties as the intra-organizational process structure as depicted in figure 2 and the Double depicted in figure 3. This result is graphically depicted in figure 6.

As a result we have a precise picture and definitions how a CNP can be extended. Indeed we recognize the concept of a **smart contract**. âĂIJA smart contract is a computerized transaction protocol that executes the terms of a contract. The general objectives of smart contract design are to satisfy common contractual conditions (such as payment terms, liens, confidentiality, and even enforcement), minimize exceptions both malicious and accidental, and minimize the need for trusted intermediariesâĂİ [1, 2]. In our model DialogDoubleExtended the reach of the smart contract is the dotted rectangle. As for MAAS services it is easy to see how the communication dialog can be modeled extending the Triple.

By executing the smart contract per transaction for as well the left hand side as the right hand side we have the correct information in the platform about what is to be received by the buyer and what is to be delivered by the seller. The same is true for the money part of the transaction between buyer and seller mediated by the platform. What is to be delivered by the seller. What has to be received by the seller. What is to be received by the buyer has to be paid by the buyer or what has been paid by the buyer has to be received by the buyer.

3.8 Causal model for eliciting requirements

A model is considered to be a representation of some object, behavior, or a system that one wants to understand [33]. In everyday life we are on a day to day basis involved in making decisions about what should a model should look like to become meaningful and therefore useful. In the end you want to make sure that people you work and communicate with understand your goals and your wants. A model is always the result and the start of a design process. "A design process is an abductive sensemaking process, a step of adopting a hypothesis as being suggested by the facts ... a form of inference, albeit inference of "best guesses" leaps [..]. *A logic of what might be.* It is not entirely accurate, âĂę it is the argument to the best explanation, the hypothesis that makes the most sense given observed phenomena or data and based on prior knowledge" [29]. The purpose of the TripleExtended depicted in figure 7 fulfills this purpose. In 3.5 and 3.6 we specified the causal model of a bilateral contract in its canonical form coined as a value net with a precise mathematical structure in terms of a dimensionalized Petri Net. In section 3.7 we extended the model with secure privacy preserving communication flows. Notice that coulouring of the Petri Net model was not necessary. As addressed in the introduction the key issue underlying a design that implements MAAS providers enabling citizens to actually buy the MAAS service is **where to start modeling** and how to express the notions in some language that is understandable for all stakeholders of the MAAS service system. The first step is to identify the business model of the platform here MAAS services as an instantiation of smart-mobility services. The business functions as a result are Buy → Search → Find → Sell. Than we reason from right to left. The result is that the value net of the platform is the first model we have to elicit that is modeling the Search → Find mechanism. Than we model the Buy → Sell mechanism. As a result we have the TripleExtended.

4 CONCLUSION

Structure defined as an assembly of components should always be studied in tandem with an associated process, whatever this process may be. The same duality seems valid for the *notion* of information or systems design. What does information do for each process, and can we find one abstract level of representation that stays away from the details of implementation [3]. In this paper we introduced the notion of value net to model MAAS as a platform as a foundation to elicit requirements in a principled way. Value net modeling offers a precise abstract representation which provides in the detailed informational requirements in a canonical form and it connects i.e. implements the abstract notion of Service Oriented Architecture characterizing systems as defined in section 3.2. without loss of crucial informational elements.

REFERENCES

[1] 2018. Essays, Papers, and Concise Tutorials. http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart.contracts.html

[2] 2018. Smartcontracts. https://archive.is/SK4vr

[3] Pieter Adriaans and Johan van Benthem. 2008. *Handbook of Philosophy of information.* Vol. 8. Elsevier B.V.

[4] Raphael Amit and Christoph Zott. 2001. Value creation in E-business. *Strategic Management Journal* 22, 6âĂŘ7 (2001), 493–520. https://doi.org/10.1002/smj.187 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.187

[5] Leonidas Antholoupos and Amel Attour. 2018. Smart transportation applications's business models: A comparison. In *WWW'18 Companion: The webconference Companion.* ACM, New York, NY, USA, 6. https://doi.org/10.1145/3184558/3191520

[6] Leonidas Anthopoupos, Marijn Janssen, and Viskanth Weekakkody. [n. d.]. A unified Smart City Model (USCM) for smart city conceptualisation and benchmarking. *International Journal of e-Government research (IJEGR)* 12, 2 ([n. d.]), 76–92.

[7] Kenneth Arrow, Amartya Sen, and Kotaro Suzumura. 2011. Chapter Thirteen - Kenneth Arrow on Social Choice Theory. In *Handbook of Social Choice and Welfare*, Kenneth J. Arrow, Amartya Sen, and Kotaro Suzumura (Eds.). Handbook of Social Choice and Welfare, Vol. 2. Elsevier, 3 – 27. https://doi.org/10.1016/S0169-7218(10)00013-4

[8] International Data Spaces Association. 2018. IDS REFERENCE ARCHITECTURE MODEL INDUSTRIAL DATA SPACE. (2018), 92.

[9] DG Bereikbaarheid. 2017. Market consultation Mobility as a Service (MaaS) in the netherlands. Public Tender. Retrieved January 2019 from https://www.tenderned.nl/tenderned-tap/aankondigingen/146777;section=2

[10] Jonathan Billington, Amar Kumar Gupta, and Guy Edward Gallasch. 2008. Modelling and Analysing the Contract Net Protocol - Extension Using Coloured Petri Nets. In *Proceedings of the 28th IFIP WG 6.1 International Conference on Formal*

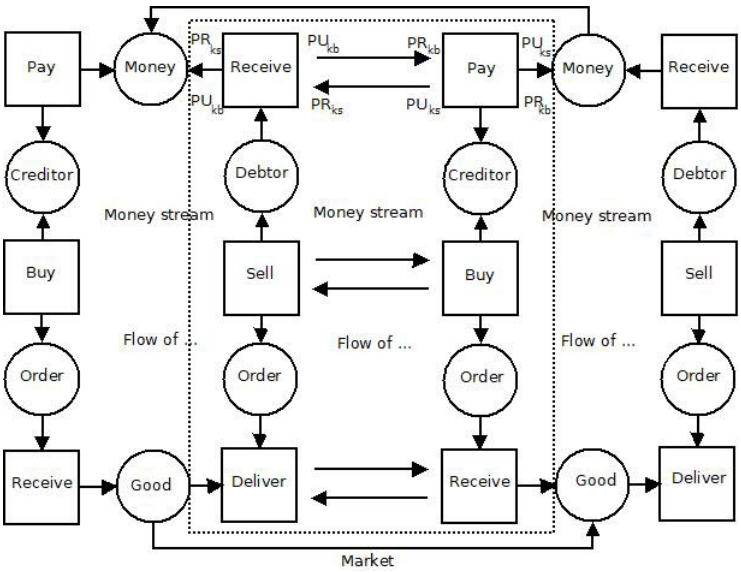


Figure 6: DialogDoubleExtended

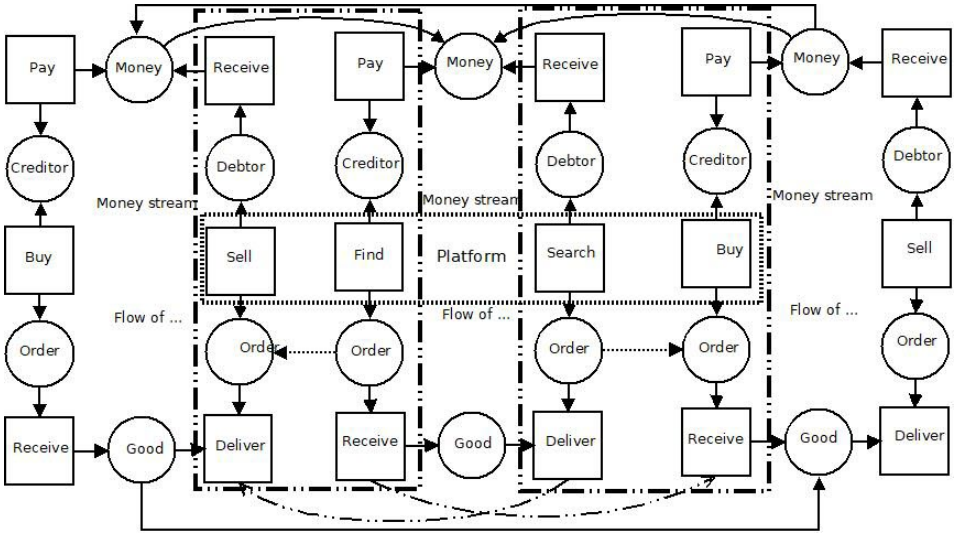


Figure 7: TripleExtended

Techniques for Networked and Distributed Systems (FORTE '08). Springer-Verlag, Berlin, Heidelberg, 169–184. https://doi.org/10.1007/978-3-540-68855-6_11

[11] G. Boella and L. van der Torre. 2004. Contracts as legal institutions in organizations of autonomous agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.* 948–955.

[12] Guido Boella and Leendert van der Torre. 2004. Regulative and Constitutive Norms in Normative Multiagent Systems. In *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning (KR'04)*. AAAI Press, 255–265. http://dl.acm.org/citation.cfm?id=3029848.3029882

[13] J. Efrim Boritz. 2005. IS practitioners views on core concepts of information integrity. *International Journal of Accounting Information Systems* 6, 4 (dec 2005), 260–279. https://doi.org/10.1016/j.accinf.2005.07.001

[14] Rob Christiaanse, Aditya Ghose, Pablo Noriega, and Munindar P. Singh. 2014. Characterizing Artificial Socio-Cognitive Technical Systems. http://digital.csic.es/handle/10261/131371

[15] Rob Christiaanse, Paul Griffioen, and Joris Hulstijn. 2015. Adaptive Normative Modelling: A Case Study in the Public-Transport Domain. In *14th Conference on e-Business, e-Services and e-Society (I3E) (Open and Big Data Management and Innovation)*, Marijn Janssen, Matti Maki, Jan Hidders, Bram Klievink, Winfried Lamersdorf, Sebastiaan van Loene, and Anneke Zuiderwijk (Eds.), Vol. LNCS-9373. Delft, Netherlands, 423–434. https://doi.org/10.1007/978-3-319-25013-7_34 Part 4: E-Business, E-Services and E-Society.

[16] Rob Christiaanse, Paul Griffioen, and Joris Hulstijn. 2015. Reliability of Electronic Evidence: An Application for Model-based Auditing. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law (ICAIL '15)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2746090.2746098

[17] Rob Christiaanse and Joris Hulstijn. 2012. Control Automation to Reduce Costs of Control. In *Advanced Information Systems Engineering Workshops*, Marko Bajec and Johann Eder (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 322–336.

[18] Rob Christiaanse and Joris Hulstijn. 2013. âĂIJControl Automation to Reduce Costs of Control.âĂİ. *International Journal of Information System Modeling and Design (IJISMD* 4, 4 (2013), 27–47.

[19] H. Coppens, T. Geel van. 2013. Urban architectural design and scientific research: how to save an arranged marriage. , 496 pages.

[20] Auriol Degbelo and Tomi Kauppinen. 2018. Increasing transparency through web maps. In *WWW'18 Companion: The webconference Companion*. ACM, New York, NY, USA, 6. https://doi.org/10.1145/3184558/3191515

[21] Hanan El Bakkali, Hamid Hatim, and Ilham Berrada. "2010". A New Task Engineering Approach for Workflow Access Control. In *Computational Intelligence in Security for Information Systems 2010*, Alvaro Herrero, Emilio Corchado, Carlos Redondo, and Angel Alonso (Eds.). "Springer Berlin Heidelberg", "Berlin, Heidelberg", "203–212".

[22] R. Foque. 2003. Research in Design Science. *ADSC* 10-11 (2003).

[23] A. Heylighen, H. Cavellin, and M Bianchin. 2009. Design in Mind. *Design Issues* 25, 1 (2009).

[24] Jingwei Huang, David M. Nicol, Rakesh Bobba, and Jun Ho Huh. 2012. A Framework Integrating Attribute-based Policies into Role-based Access Control. In *Proceedings of the 17th ACM Symposium on Access Control Models and Technologies (SACMAT '12)*. ACM, New York, NY, USA, 187–196. https://doi.org/10.1145/2295136.2295170

[25] L. Hurwicz. 1960. *Optimality and informational efficiency in resource allocation processes*. Stanford Universiy Press.

[26] L. Hurwicz and S. Reiter. 2006. *Designing economic mechanisms*. Cambridge University Press, New York. 341 pages.

[27] ITS. [n. d.]. Smarttrafic. Retrieved Jan 15, 2019 from http://www.its-finland.fi/index.php/en/

[28] M. Kamargianni and M. Mathyas. 2017. âĂIJThe business ecosystem of Mobility-as-a-ServiceâĂİ. (2017), 14. 96th Transportation Board Annual meeting.

[29] J. Kolko. 2010. Abductive Thinking and Sensemaking: The drivers of Design Synthesis. *Design Issues* 26, 1 (2010), 14.

[30] William Kuechler and Vijay Vaishnavi. 2012. A Framework for Theory Development in Design Science Research: Multiple Perspective. *Journal of the Association for Information System* 13, 6 (2012).

[31] David Lewis. 2002. *Conventions*. Blackwell.

[32] Satoshi Nakamoto. 2008. Bitcoin: A Peer-to-Peer Electronic Cash System. , 9 pages.

[33] D.A. Norman. 1993. *Things that make us smart*. Basic Books. 281 pages.

[34] OASIS. 2012. Reference Architecture Foundation for service oriented architecture, version 1. (2012), 118.

[35] Jean-Charles Rochet and Jean Tirole. 2004. Two-sided markets: An overview. (2004), 44.

[36] Bruce Schneier. 1996. *Applied Cryptography. Protocols, algorithms and source code in C* (second ed.). Katherine Schowalter. 675 pages.

[37] Kalina Staykova and Jan Daamsgaard. 2015. A typology of multi-sided platforms: the core and the periphery. 23rd European Conference on Information Systems (ECIS), Munster, 16.

[38] Hideaki Takeda, Paul Veerkamp, Tetsuo Tomiyama, and Hiroyuki Yoshikawa. 1990. Modeling Design Processes. *AI Mag.* 11, 4 (1990), 12. http://dl.acm.org/citation.cfm?id=95788.95795

[39] T. R. Weil and E. Coyne. 2013. ABAC and RBAC: Scalable, Flexible, and Auditable Access Management. *IT Professional* 15 (05 2013), 14–16. https://doi.org/10.1109/MITP.2013.37

[40] Oliver E. Williamson. 1973. Markets and Hierarchies: Some Elementary Considerations. *The American Economic Review* 63, 2 (1973), 316–325.

[41] Haiping Xu and S. M. Shatz. 2001. An agent-based Petri net model with application to seller/buyer design in electronic commerce. In *Proceedings 5th International Symposium on Autonomous Decentralized Systems*. 11–18. https://doi.org/10.1109/ISADS.2001.917390

[42] H. Zimmerman. 1980. RapporteurâĂŹs Group on âĂIJLayered Model for Public Data Network. *IEEE Transactions on Communications* 28, 4 (1980), 8.

[43] Christoph Zott and Raphael Amit. 2013. The business model: A theoretically anchored robust construct for strategic analysis. *Strategic Organization* 11, 4 (2013), 403–411. https://doi.org/10.1177/1476127013510466 arXiv:https://doi.org/10.1177/1476127013510466

Supporting Urban Air Mobility with Citizen Participatory Noise Sensing: A Concept

Hinnerk Eißfeldt
Department of Aviation and Space Psychology
DLR German Aerospace Center, Hamburg, Germany, hinnerk.eissfeldt@dlr.de

ABSTRACT

Looking ahead, in about 20 years, there is likely to be urban air mobility in larger cities across the globe. If economic predictions come true, thousands of air taxi flights will take place daily in capital cities – not only in megacities. Noise generated by urban flight mobility has been identified as a critical factor in this development. A concept is proposed to help raise the tolerance level for urban air noise among communities as well as of individual residents by means of transparency. This concept views residents as stakeholders in urban air mobility and widens the call for continuous noise measurements of vertical take-off and landing operations on individual site basis [1] by residents voluntary on-site data collection enabled by smartphone-based participatory noise sensing (PNS). In the presentation of this concept, this discussion paper describes important aspects of social acceptance of urban air mobility.

CCS CONCEPTS

• Applied computing~Aerospace • Networks~Public Internet • Human-centered computing~Collaborative content creation • Hardware~Sound-based input / output

KEYWORDS

Urban air mobility; environmental noise; air taxi; smartphone; community map; participatory noise sensing; noise annoyance

ACM Reference format:

Hinnerk Eißfeldt. 2019. Supporting Urban Air Mobility with Citizen Participatory Noise Sensing: A Concept. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3317059

1 Introduction

Looking ahead, in about 20 years, there is likely to be urban air mobility (UAM) in larger cities across the globe. Air taxis will serve individual on-demand transportation needs. They will also connect important transportation nodes, for instance the city center with the airport, and will provide fast journeys between train stations or across rivers and lakes. They are envisaged to operate around-the-clock. Droneports (also called Vertiports or Skyports) will develop at various points in the city, allowing easy access, transfer to other modes of transportation, and technical support such as battery charging and maintenance for air taxis. This, together with existing no-fly zones due to safety restrictions, will lead to the development of certain routes above the city. A typical approach for aviation administrations is to bundle such routes in corridors to keep drone traffic separated from other airspace users (for a recent example see [2]). Working or living under or near such corridor will overly expose citizens to air taxi noise emissions and could become a limiting factor for urban air mobility. This discussion paper argues that citizens have to be viewed as stakeholders in urban air transportation, regardless of whether they intend to use it or not, and that a concept of resident PNS will be beneficial to UAM. Web-based services and smartphones allow the current state of information to be easily accessed and updated, thus enabling them to be used to foster urban air mobility.

2 Noise Concerns

Organizing UAM in a manner that is similar to today's commercial air traffic, i.e. using certain corridors (or tubes, see [3]), is meant to guarantee safe operations and keep different kinds of air traffic apart from each other. However, in doing so the noise problem will not only exist in the vicinity of droneports, where vehicles take off and land, but also for citizens living or working below these corridors, which are projected to be at an altitude between 500 and 1500 feet. "One potential outcome of scaled-up drone operations is an increase in urban noise volume exceedances above legal or desired limits" [4]. An expected decrease in quality of life and/or real estate values due to urban air mobility can lead to lawsuits and/or restrictions. Strong public opposition could finally become a show-stopper for urban air mobility, at least in western megacities, which is also the reason why the aviation industry sees "involving community engagement now vitally important" [5].

Among all sources of environmental noise, research has shown that the noise generated from aviation is the most annoying [6]. Own research has shown that, among all concerns, those relating to noise are the strongest predictor of acceptance of civil drones: people who are concerned about noise show least acceptance. Furthermore the usage of air taxis finds only little acceptance among the public. In a representative study in

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317059

Germany, 85% of the respondents in a telephone survey indicated they would not use air taxis [7]. Even among those who would consider using UAM, their acceptance is likely to vanish when air taxis are flying above their house or workplace.

Annoyance can be seen as a multidimensional construct comprising three elements: firstly the experience of repeated noise-related disturbance as well as the behavioral response to cope with it, secondly the emotional/attitudinal response to the sound and its disturbing impact, and lastly the perceived lack of capacity to cope with the noise [8]. The coping style can be either problem-oriented or emotion-oriented, whereas the problem-oriented coping is more proactive and the latter more reactive in nature.

Communities can differ in their tolerance to noise. The community noise tolerance level (CTL) is described as the noise level at which half of the population reports being "highly annoyed" and is measured using a standard set of questions. Research has found that the same objective noise level measured in dB can lead to different CTLs at different locations. In addition to personal, situational, and cultural influences, the rate of abrupt change in noise levels during recent years has been identified as factor that influences the community noise level, with airports that have a high rate of change leading to higher annoyance levels than airports with a low rate of change [9].

Most of the findings on aviation noise relate to classic air traffic, to jet aircraft, and airports, so they have to be downscaled when applied to UAM. The industry seems confident in the potential to produce low-noise vehicles [1]. However, with 24/7 operation envisaged, with predicted airspeeds exceeding 100 mph, and with the basic physical fact that when air meets a structure it produces sound (true for propellers as well for the cabin), the noise problem will be persistent with UAM. Not only that UAM noise will be an additional factor to other environmental noise sources such as road, rail, aviation, and industry; It will "come on top" for all citizens living or working on higher floors in the tall buildings of our urban canyons.

3 Participatory noise sensing

Noise mapping using smartphones allows citizens to actively engage and contribute data of personal relevance [10]. Real-time measurements with mobile applications represent the individual's perception of noise. Using individual smartphones for sensing might face calibration issues, however, every new generation of devices outperforms its predecessors. Noise mapping using smartphones has made large advances in recent years. Applications available today enable residents to submit noise data and to provide additional information directly from their smartphone [11]. When proposing a concept of citizen PNS, Guillaume and colleagues [12] list more than ten different smartphone tools for noise mapping. An in-depth description of "NoiseCapture," an open-science crowdsourcing application for producing community noise maps is provided by Picaut et al. [13]. NoiseCapture [14] is a smartphone application that is freely available and was developed in the framework of a research network partially funded by the European Community. The application is dedicated to the android platform and has been in use since August 2017. Using such a tool, residents can contribute to noise mapping with data specifically relevant to them while still maintaining user anonymity.

4 Proposed urban air mobility management

An urban air traffic management (ATM) system is installed at the city level for public governance and tracks all flights in urban airspace. For noise monitoring, the system calculates flight noise-profiles either by using emission characteristics from a repository of all air-taxi models and combining it with relevant data on current weather conditions (e.g. wind, cloud base, humidity) or by using real-time measurements provided by stationary devices or mobile applications. Onsite measurements are superior to other methods as they are closest to the true noise perceived by the citizen. Artificially calculated measures can never include all relevant data and are likely to underestimate the true noise values. Measurement data from unsupervised stations or data harvested by volunteers will undergo sound spectrum analysis to be checked for outliers and faithfulness before being included in the database. Combining UAM traffic data with incoming noise measurement files using timestamps and/or geopositioning data will allow UAM noise to be distinguished from other environmental noise sources when necessary.

Noise profiles will be generated and analyzed under public control. The data will be stored geographically like footprints or "noise tracks" on regional maps. Track colors represent noise intensity at ground level, with track density showing the frequency of flights at a given location. This map is constantly updated at weekly intervals and provides the ongoing base for the calculation of noise impact. It is accessible for the public.

Whenever possible, the urban air traffic management system is able to adjust requested flight patterns so that the noise-burden is shared in an objective and transparent manner. For instance within a dynamic skylane network, altitudes could be adjusted or the lateral route of flight can be adapted based on prior noise load at a given location.

5 Conclusions

Air mobility will come to our cities within the next decades. It will come as an additional platform in urban mobility and it will be an additional source of environmental noise for citizens. It will receive extra public attention because it will be used only by few people, but it will produce elevated noise volumes for many. One way to support the successful introduction of UAM is to invite residents to actively participate in noise sensing by using smartphone applications. Such application should be available at no cost for the user and for all platforms. It could be even more successful if additional smart city functions were offered, such as linking with other transport systems, public services, or an option to request a temporary flight ban over a certain location, e.g. when there is a neighborhood festivity.

Making reference to individual noise loads, as proposed in this concept, would support the onset of UAM while safeguarding the wellbeing of citizens. The less noisy urban air

taxi operation turns out to be for the residents, the more successful its introduction can be – at little extra cost.

REFERENCES

[1] Uber 2016. Fast-forwarding to a future of on-demand urban air transportation. Uber Elevate.

[2] Ministry of Civil Aviation of the Government of India (2019). Drone ecosystem policy roadmap.

[3] Balakrishnan, K., Polastre, J., Mooberry, J., Golding, R. & Sachs, P. (2018). Blueprint for the sky. The roadmap for the safe integration of autonomous aircraft. Airbus A3.

[4] International Transport Forum (OECD) (2018). (Un)certain Skies? Drones in the World of Tomorrow. Retrieved from https://www.itf-oecd.org/uncertain-skies-drones [29 Aug 2018]

[5] Civil Air Navigation Services Organisation (2015). Managing the impacts of aviation noise. A guide for airport operators and air navigation service providers. CANSO.

[6] Guski, R., Schreckenberg, D. & Schuemer, R. (2017). WHO Environmental Noise Guidelines for the European Region: A Systematic Review on Environmental Noise and Annoyance. International Journal of Environmental Research and Public Health 14(12):1539. DOI: 10.3390/ijerph14121539.

[7] Eißfeldt, H., Vogelpohl, V., Stolz, M., Papenfuß, A., Biella, M., Belz, J. & Kügler, D. (2018) The acceptance of civil drones in Germany. Manuscript submitted for publication.

[8] Schreckenberg, D., Belke, C. & Spilski, J. (2018). The Development of a Multiple-Item Annoyance Scale (MIAS) for transportation noise annoyance. Int.J. Environ. Res. Public Health 2018, 15, 971.

[9] Gelderblom , F. B., Gjestland, T., Fidell, S. & Berry, B.(2017). On the stability of community tolerance for aircraft noise. AAA Vol.103, 17-27.

[10] Maisonneuve, N., Stevens, M. & Ochab, B. (2010). Participatory noise pollution monitoring using mobile phones. Information polity, 15, 51-71.

[11] Becker, M., Caminiti, S., Fiorella, D., Francis, L. Gravino, P. et al. (2013). Awareness and learning in participatory noise sensing. PloS ONE 8(12): e81638. DOI: 10.1371/journal.pone.0081638.

[12] Guillaume, G., Can, A, Petit, G., Fortin, N., Palominos, S., Gauvreau, B., Bocher, E. & Picaut, J. (2016). Noise mapping based on participative measurements. Noise Mapping, 2016; 3:140-156

[13] Picaut, J., Fortin, N., Bocher, E., Petit, G., Aumond, P. & Guillaume, G. (2019). An open-science crowdsourcing approach for producing community noise maps using smartphones. Building and Environment. Vol. 148, pp. 20-33.

[14] https://www.ifsttar.fr/en/online-resources/science-and-society/spatial-planning/focus-on/an-application-to-evaluate-our-sound-environment-noisecapture/ Accessed on 12 Jan 2019.

Integrating Route Attractiveness Attributes into Tourist Trip Recommendations

Daniel Herzog
Department of Informatics
Technical University of Munich
85748 Garching bei München, Germany
herzogd@in.tum.de

Sherjeel Sikander
Department of Informatics
Technical University of Munich
85748 Garching bei München, Germany
sherjeel.sikander@tum.de

Wolfgang Wörndl
Department of Informatics
Technical University of Munich
85748 Garching bei München, Germany
woerndl@in.tum.de

ABSTRACT

Tourist trip recommender systems (RSs) support travelers in identifying the most attractive points of interest (POIs) and combine the POIs along a route for single- or multi-day trips. Most RSs consider only the quality of POIs when searching for the best recommendation. In this work, we introduce a novel approach that also considers the attractiveness of the routes between POIs. For this purpose, we identify a list of important attributes of route attractiveness and explain how to implement our approach using three exemplary attributes. We develop a web application for demonstration purposes and apply it in a small preliminary user study with 16 participants. The results show that the integration of route attractiveness attributes makes most people choose the more attractive route over the shortest path between two POIs. This paper highlights how tourist trip RSs can support smart tourism. Our work aims to encourage further discussion on collecting and providing environmental data in cities to enable such applications.

CCS CONCEPTS

• **Information systems → Web applications**; **Recommender systems**.

KEYWORDS

Recommender System, Tourist Trip Design Problem, Route Attractiveness, Environmental Data, User Study

ACM Reference Format:
Daniel Herzog, Sherjeel Sikander, and Wolfgang Wörndl. 2019. Integrating Route Attractiveness Attributes into Tourist Trip Recommendations. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317052

1 INTRODUCTION

Generally, tourists exploring a city wish to visit as many points of interest (POIs) as possible. However, visiting all POIs is usually not feasible because of constraints, such as the time available for the trip. The problem of designing tourist trips covering multiple POIs along enjoyable routes is called the Tourist Trip Design Problem (TTDP) [22]. In its simplest specification, the TTDP is identical to the Orienteering Problem (OP), a combinatorial optimization problem in which several locations with an associated profit have to be visited within a given time limit. Each location may be visited only once, and the aim is to maximize the overall profit gained on the tour [19]. Practical applications of the TTDP can be integrated in recommender systems (RSs) which help users to identify the most attractive POIs in an area of interest and combine them to devise single- or multi-day trips.

In most tourist trip algorithms and applications, the total profit of a trip is the sum of the profits of the locations. However, we argue that tourists do not always want to take the shortest route between two POIs. Rather, the perceived quality of a tourist trip depends also on the attractiveness of the routes between the POIs. For example, a trip becomes more attractive when the route between two POIs is a relaxing walk in a green area instead of a walk by a loud street, even if this means taking a detour. Furthermore, when too many tourists take the same routes, the recommendations can be adapted to better balance the tourist flows in the city, thereby helping reduce the crowd and pollution in these areas. Consequently, tourist trip RSs should be able to determine the profits of routes by considering all relevant route attractiveness attributes and to adapt the recommendations accordingly.

In this paper, we present a solution to integrate route attractiveness attributes into a tourist trip RSs for walking tourists. Our approach allows the integration of any data influencing the quality of routes and adjusting the importance of each attribute according to the user's needs. We illustrate our approach by explaining the integration of three attributes as an example. In addition, we present a web-based application implementing our approach and evaluate the recommended routes in a user study. A more detailed description of our approach can be found in [16]. The advantages of our system are twofold. Tourists receive recommendations for personalized trips tailored to their needs. Further, cities and tourist spots can promote the development of smart tourism by providing the users with all relevant information about the attractions to visit, how to reach them, and any form of environmental conditions influencing their trip. Hence, our main contribution is an innovative, practical smart tourism application that promises benefits with regard to the economy, environment and mobility in cities. The aim of this paper is to encourage discussion on the environmental data that should be collected in cities to enable these benefits, how to collect these data, and how to make the data accessible to promising applications.

The remainder of this paper is organized as follows: In Section 2 we summarize relevant related work. We define route attractiveness

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317052

and identify relevant attributes for a tourist trip RS in Section 3. In Section, 4 we explain the edge weight calculation of selected attributes. We present the implementation of our approach and our web-based application in Section 5. The preliminary user study we conducted and its results are presented in Section 6. Finally, Section 7 presents the conclusions and suggestions for future work.

2 RELATED WORK

RSs support tourists in finding best travel destinations, POIs, and travel packages combined of multiple travel-related items [1]. Practical applications that solve the TTDP have been developed in the recent years. The City Trip Planner, introduced by Vansteenwegen et al. [20], was the first practical application to solve the TTDP. It provides recommendations for single- or multi-day trips and can include lunch breaks into the recommendation. CompRec-Trip is a RS for sets or sequences of POIs; it has a graphical user interface allowing users to customize the recommended trips [25]. Gavalas et al. [6] developed DailyTRIP, a web and mobile web application that takes into account the opening days and hours. Recently, Herzog et al. [9] presented TourRec, a mobile RS that provides suggestions for personalized trips to individuals and groups.

All of these RSs solve the TTDP by maximizing the sum of POI profits while trying to minimize the distances between the POIs. Only few works have considered the quality of routes between POIs when determining the total value of a trip. Souffriau et al. [17] introduced the Arc Orienteering Problem (AOP), a variant of the OP. The AOP assigns a cost and a non-negative profit to routes (*arcs*) instead of locations. Souffriau et al. presented a greedy randomized adaptive search procedure to solve the AOP and implemented it in a cycle route planner to recommend pleasant routes for cyclists. Verbeeck et al. [23] presented the Cycle Trip Planning Problem (CTPP), an extension of the AOP. In the CTPP, a vertex can be visited twice. Vansteenwegen et al. [21] introduced another extension of the AOP, namely, the Mixed OP; in this problem, scores are assigned to routes as well as locations. Gavalas et al. [5] worked on recommending scenic routes, i.e., walking paths of tourist value. They integrated their approach in a context-aware mobile city guide for Athens, Greece with 18 scenic routes. A user study confirmed that their approach can recommend tours of high quality. Querica et al. [14] introduced a different approach to recommend trips. Their recommendations suggest routes that are perceived as pleasant, thereby trying to maximize the perceived happiness of users while adding only a few extra walking minutes to the trip.

However, to the best of our knowledge, there is no tourist trip RS that considers all types of route attractiveness attributes, such as air pollution, greenery, and street quality, to find the optimal trip. In this paper, we explain how such attributes can be integrated into TTDP algorithms. The novelty of our approach is its flexibility that allows easy addition or removal of attributes and adaption of the route attractiveness calculation to the recommendation's context and the user's needs.

3 ROUTE ATTRACTIVENESS

The influence of environmental attributes on people's decisions to participate in outdoor activities, such as walking or cycling, has been explored earlier. We analyzed previous studies [8], literature reviews [11, 18], developed frameworks [12], and models [2, 15] that dealt with the influence of environmental attributes on the choice of walking routes. It is not possible to directly compare the results of all works as they used different methodologies or were limited to specific target groups, such as elderly people. Based on our findings, we thus devised a novel, subjective list of route attractiveness attributes that should be considered when evaluating routes in a tourist trip RS.

Table 1 lists all route attractiveness attributes that we identified and their impact on the quality of a walking route. The attributes are classified into three categories: attributes that are likely to affect the route, attributes that are somewhat likely to affect the route, and attributes that are less likely to affect the route or for which the data are insufficient to determine the probability. The listed attributes are not ordered.

Table 1: Route attractiveness attributes for tourist trip RSs. The (+) sign following an attribute's name indicates a positive impact on a walking route, and the (−) sign, a negative impact.

Attribute	Probability
Aesthetics: Trees (+)	High
Aesthetics: Pollution (−)	High
Aesthetics: Cleanliness (+)	High
Permeability: Pavements (+)	High
Traffic: Speed (−)	High
Walking Surface: Maintenance (+)	High
Personal Safety: Surveillance (+)	High
Aesthetics: Landscaping (+)	Medium
Permeability: Intersection Distance (−)	Medium
Traffic: Volume (−)	Medium
Personal Safety: Lightening (+)	Medium
Traffic: Crossings (+)	Medium
Traffic: Crossing Aids (+)	Medium
Streets: Width (+)	Medium
Aesthetics: Parks (+)	Medium
Permeability: Slopes (−)	Medium
Permeability: Stairs (−)	Medium
Traffic Control Devices (+)	Low / Unclear
Walking Surface: Continuity (+)	Low / Unclear
Traffic: Verge Width (+)	Low / Unclear
Destination: Shops (+)	Low / Unclear
Traffic: Noise (−)	Low / Unclear
Facilities: Places to rest (+)	Low / Unclear
Environment: Walking Trails (+)	Low / Unclear
Facilities: Shops (+)	Low / Unclear
Personal Safety: Blind Walls (−)	Low / Unclear
Aesthetics: Green strips (+)	Low / Unclear

All the earlier studies that we reviewed agree that aesthetics attributes are the most important route attractiveness attributes. These attributes include the number of trees along the path, pollution, and cleanliness. Traffic speed is an important safety hazard influencing the attractiveness of a route while the impact of traffic noise is unclear. Other important route attractiveness attributes are

the presence of pavements, maintenance of the walking surface, and personal safety in form of surveillance (people around, avoiding empty streets, etc.).

We suggest that smart tourism applications, such as tourist trip RSs, should consider all attributes in the first category and examine whether attributes from the secondary category should be considered before recommending routes to people. Further research is necessary to evaluate the impact of the attributes from the last category, and their impact may also highly depend on the use case.

4 EDGE WEIGHT CALCULATION OF EXEMPLARY ATTRIBUTES

Our approach recommends tourist trips by running a shortest path algorithm on a graph of vertices and edges. However, our approach does not use only the distance between two vertices to determine the edge weight (*cost*). It also considers the presence or absence of the relevant route attractiveness attributes listed in Table 1.

In this section, we explain the calculation of the edge weights in such a shortest path algorithm for three examples: trees, pollution, and cleanliness. These are the most relevant attributes for our scenario. We then present the final edge weight calculation taking into account all three attributes. Our goal is to present a flexible and extendable solution for integrating attractiveness attributes. Providers of tourist trip RSs should be able to easily add or replace attributes, depending on the available data sources. Users should be able to individually adjust the importance of each attractiveness attribute while exploring a city.

4.1 Aesthetics: Trees

The tree edge weight is based on the tree density on a route. The greater the number of trees on the edge, the higher is the tree density and vice versa. A higher tree density corresponds to a lower edge weight in the graph.

In order to weight the different tree densities, we utilize a weighting method described by Giles-Corti et al. [7]. For our purpose, we normalize the weights to a scale from 0 to 5 (Table 2).

Table 2: Weights for tree density according to [7] and after normalization (exact values in brackets).

Tree Density	Weight	Normalized Weight
Many trees touching	14.3	5
Some trees touching	11.4	4 (3.986)
Trees close but do not touch	8.6	3 (3.007)
Trees spread apart	5.7	2 (1.993)
Sparse trees	2.86	1
No trees	0	0

We use the horizontal spread of the tree when viewed from the top, i.e., the crown spread, to define tree density. For the sake of simplicity, we use a generic crown radius of 5 m for every tree in our algorithm. This is roughly the average crown size of the *Tilia cordata* Mill. species, which is the species of trees most commonly planted in Berlin, Germany (35 %) [13]. We use the following function to estimate tree density:

$$DensityScore(Edge_i) = \frac{EdgeLength(Edge_i)}{NumberOfTrees(Edge_i)}, \quad (1)$$

where $Edge_i$ is the ith edge in the graph; $EdgeLength$ is the distance of the edge in meters; and $NumberOfTrees$ is the total number of trees assigned to the edge. Using this equation and assuming a generic crowd radius of 5 m, we estimated tree density as specified in Table 3.

Table 3: Tree density categorization based on density score.

Tree Density	Density Score (x)
Many trees touching	$x < 5$
Some trees touching	$5 \leq x < 15$
Trees close but do not touch	$15 \leq x < 25$
Trees spread apart	$25 \leq x < 35$
Sparse trees	$35 \leq x$
No trees	0

The tree edge weight is eventually calculated by dividing the density score by the normalized weight of the respective tree density category in Table 2. When there are no tress, the edge length is divided by 0.5.

4.2 Aesthetics: Pollution

Air quality is measured differently by different countries. Pollutants are measured over a certain defined period, and the density of the pollution is usually measured in micrograms per cubic meter. Some of the most common air pollutants are nitrogen dioxide (NO2), ozone (O3), fine particulate matter (PM2.5), and coarse particulate matter (PM10). In order to standardize the concentration values of different pollutants, we use the Common Air Quality Index (CAQI) (Table 4) [3]. The CAQI is an index that compares air quality across different European countries. It has been used in the data on the website www.airqualitynow.eu since 2006.

Table 4: CAQI values corresponding to different pollutant concentrations [3].

Index Class	Grid	Pollutant (hourly) density in μ/m^3			
		NO2	PM10	O3	PM2.5 (opt.)
Very high	>100	>400	>180	>240	>110
High	75–100	200–400	90–180	180–240	55–110
Medium	50–75	100–200	50–90	120–180	30–55
Low	25–50	50–100	25–50	60–120	15–30
Very low	0–25	0–50	0–25	0–60	0–15

If two or more pollutants have different CAQI values for a region, the higher value is considered to be the overall CAQI value as the worst grid value determines the overall index class.

We use the CAQI values to map air pollution values to edges. The air pollution weight is calculated by dividing the CAQI value of the edge by 25 and multiplying it with the edge length. For very low air pollution CAQI values, we divide the CAQI value of the edge by 50 to reduce the costs of edges with very low pollution even more strongly.

4.3 Aesthetics: Cleanliness

There is no standard way of measuring cleanliness as it greatly depends on the public perception of littering. In this work, we measure cleanliness as littering on a scale of 0 to 10. The littering value could be reported by citizens through mobile applications, for example. Values less than 5 are considered as low or medium littering while values greater than or equal to 5 are considered as high littering. The value of 0 represents either no littering or unavailability of data.

The littering weight for high littering is calculated by multiplying the littering value by the edge length. For low littering values, we divide the littering value by 2 before multiplying it by the edge length. For no littering or no data, we multiply the edge length by 0.1.

4.4 Final Edge Weight

The final edge weight e takes the edge length and the attractiveness attributes into account. Our proposed calculation assigns weights to the attributes:

$$e = (x * treesWeight) + (y * airpollutionWeight) + (z * litterWeight), \quad (2)$$

where x, y, and z determine the importance of each attractiveness attribute. They can either be fixed, provided at runtime by the user, calculated from user preferences, or learned through user behavior. The dynamic nature of these values makes the algorithm adjustable, allowing us to meet the requirement of a flexible and extendable solution to integrate attractiveness attributes. For example, if the presence of trees is more important than no littering for a user, x should be greater than z. If the user wants to completely ignore an attribute, the corresponding value is set to 0. Furthermore, additional attributes with the desired weight can be added at any time. In this work, we initially set x, y, and z to 1. We evaluated the importance of each attribute, as described in Section 6.1.

5 IMPLEMENTATION

The calculated final edge weight can be interpreted as the cost for traveling between two POIs in a tourist trip RS. Shortest path algorithms, such as algorithms based on Dijkstra's algorithm, can be used to incorporate the costs when recommending tourist trips [24]. In this case, the RS first identifies the POIs that the user should visit during a trip. Then, the routing algorithm finds the best routes between the POIs, considering the calculated costs for each route.

In order to execute such an algorithm, we need map data to build a graph of vertices and edges. In the following text, we explain the pre-processing of map data and our example attributes, and the mapping of these attributes to the graph. For our prototype, we used the open source vector map data source OpenStreetMap to access to the required data. We implemented our approach using only a small extract of OpenStreetMaps since loading the entire map data for a large city or region and performing route operations on it is a memory-intensive task. In this work, we used an extract of the city center of Munich, Germany.

In OpenStreetMap, ways are essentially a collection of connected nodes. We can consider ways as edges for simplicity of discussion. A node in OpenStreetMap can be a street intersection, a bench, or any other point specific information. Nodes, such as benches, trees, or ways that are only used to define boundaries of a park, for example, are not connected to other nodes, and hence, they do not specify streets or footpaths. We need to extract only those nodes that are surely connected. We use these ways from our OpenStreetMap extract to eventually build a graph. Given the latitude and longitude information of two neighboring nodes, we can calculate the distance between them to create a weighted graph that is required in our path algorithm.

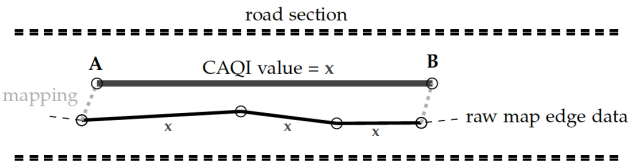


Figure 1: Mapping a CAQI vector to edges.

In the next step, we map the attractiveness attributes to the graph edges so that they can be used in the routing algorithm.

Tree data are available in OpenStreetMap files and are represented by latitudes and longitudes. We assign every tree to the edges of the nearest node. This approach increases the actual number of trees assigned to the graph; however, this tree count is only used to estimate the tree density for an edge.

Air pollution data are represented by a vector line with a start point and an end point. Since we do not have access to real pollution data, we assume the values in this work. First, we find two nodes: the nearest node for the start point and that for the end point of the CAQI vector. Within the length of a single CAQI vector, multiple edges of the graph might be present. The CAQI value needs to be translated to all the edges lying between two points A and B. Hence, we use a shortest path calculation between two nearest nodes to find all such edges. This approach is illustrated in Figure 1. Nodes A and B represent the starting and end points of a CAQI vector. They are mapped to the nearest nodes, and the shortest path between these nodes has three edges. If multiple CAQI values are mapped to the same edge, the highest CAQI value is assigned.

The litter data are also represented as a point; however, the impact is considered to be within a certain radius. Since we do not have access to real littering data, we assume their values in this work. First, we find the nearest node for every litter position. Then, we determine all the edges that are within or intersect with a 10 m radius around this nearest node. For each edge, we add the current litter value to the total litter value of this edge. It is important to note that multiple littering spots near the same edge add up, and hence the final littering value of an edge can be greater than 10.

After assigning all the weights to the edges, the final graph can be created and used by a tourist trip RS or any type of routing application to find routes taking into account the integrated attractiveness attributes. We developed a web application incorporating the aforementioned attractiveness attributes to visualize our approach and the routes that can be generated. The applications is developed using ReactJS, a javascript library for building web applications. Figure 2 shows the recommendation made by

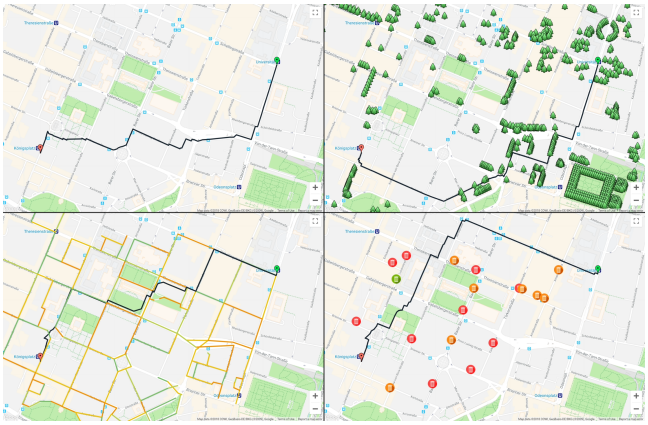


Figure 2: Visualization of route attractiveness attributes in our application and the recommended routes between two POIs. Top left: shortest path, top right: trees, bottom left: air pollution, and bottom right: cleanliness (littering).

our application regarding the shortest path between two POIs and different alternatives depending on the considered attribute.

6 PRELIMINARY USER STUDY

The main goal of the preliminary user study is to determine the impact of route attractiveness attributes on the user's decision of choosing an alternative walking route between two POIs instead of the shortest path. The user study is divided into two parts:

(1) Determination of the impact of the three considered attributes on route attractiveness.
(2) Evaluation of the recommended routes, including the route attractiveness attributes of the examples.

In all, 16 users participated in the user study. Most of the participants were students from Munich in the age range 22 and 29 years.

6.1 Attributes Questionnaire

A relevant route attractiveness attribute either has a positive impact or a negative impact on route attractiveness. We asked the participants about their decision to walk a route between two POIs if an attribute is (a) present and (b) absent.

Table 5 summarizes the results of our questionnaire. Air pollution is obviously a very critical attribute for travelers. In fact, 100% of our participants claimed that they will avoid routes with little air pollution, while the absence of air pollution is the reason for choosing an alternative route over the shortest path. A similar, less unanimous behavior can be observed when asking participants about littering. Greenery, however, seems to be a bonus for travelers. Having no greenery (e.g., no trees) does not influence their decision of choosing an alternative route.

We conclude that air pollution and littering are two dominant factors influencing the choice of a route. Hence, they should receive very high weight in a tourist trip RS. Missing greenery does not necessarily decrease the recommendation probability of a route if the route is clean. However, almost a fifth of travelers prefer routes

Table 5: Influence of the presence or absence of route attractiveness attributes on travel decisions (highest value marked in bold).

	Travel	Avoid	No influence
No air pollution	**100%**	0%	0%
Little air pollution	0%	**100%**	0%
No greenary	12.5%	18.75%	**68.75%**
Greenary	**93.75%**	0%	6.25%
No littering	**87.5%**	0%	12.5%
Littering	6.25%	**81.25%**	12.5%

with greenery when traveling. Greenery can therefore be used to choose between different routes with similar pollution. This also validates our assumption that users should be able to adjust the weights, if necessary.

6.2 Evaluation of the Recommended Trips

The 16 participants of our user study were then asked to use the application. Each participant had to specify two POIs, a starting and ending point, five times. For each POI pair, the shortest path and a route for each of the four test conditions (tree attribute only, air pollution attribute only, litter attribute only, and all the attributes combined) were calculated. For each test condition, the participants saw the attributes visualized on the application's map, as illustrated in Figure 2. Since some of our data were assumed, the participants were asked to assume that all data are real.

After every recommendation, the participants were given a comparison of the time taken in minutes for the detour compared to that for the shortest path and were asked if they would prefer the recommended route over the shortest path.

Table 6 summarizes the results. For each test condition, we conducted a binomial test to find out if the integration of route attractiveness attributes had a significant effect on the users' decision. Results show that integrating trees, air pollution, and all the attributes taken together prompt tourists to significantly more often select the recommended route over the shortest path. Less littering, however, does not seem to make people choose the recommended route more often than the shortest path.

Table 6: Ratio of users choosing the more attractive route over the shortest path (Note: *p<0.05; **p<0.01; ***p<0.001).

Algorithm	Average	Significant
Air pollution only	0.7	Yes ***
Trees only	0.76	Yes ***
Littering only	0.59	No
Combined	0.66	Yes **

7 DISCUSSION AND CONCLUSION

Previous research showed that many environmental factors have a significant influence on the perceived attractiveness of routes. The presence or absence of some of these factors is prompting people to choose routes with the given characteristics even at the cost of taking a detour. For tourists, in particular, route attractiveness

plays an important role when planning city trips. Many tourists do not want to only visit as many POIs as possible. The walking time between two locations is also part of the pleasure and can increase a traveler's happiness. Hence, attractive routes are often preferred over the shortest path if they promise a more pleasant journey. RSs for tourists should incorporate route attractiveness attributes to better adapt their recommendations to the users' individual needs.

We presented a list of attributes that influence people's decisions on the choice of walking routes. We demonstrated the integration of three such attributes into a tourist trip RS. In our preliminary user study, a significant majority chose recommended routes considering attractiveness attributes over the shortest path. Future work should provide solutions to integrate all attributes with a large impact on a walking route's quality. A big challenge for the developers of smart tourism and mobility applications is the availability of such data. In this work, we were only able to get real tree data for our test area. For cities, data have to be collected by deploying sensors to measure air pollution, for example. Collecting air pollution data for single routes or tourist spots is a challenging task as air quality is usually monitored using networks of static and sparse measurement stations [10]. More fine-grained measurements can be achieved by involving citizens who use air quality monitoring kits to measure pollution at home [4], or mobile applications to report problems, such as road damages and lack of accessibility. The data should be made public and accessible in real-time to allow developers of smart city applications to integrate them into their services.

We presented one example of an eco-sensitive, smart tourism application, i.e., a tourist trip RS. Route recommendations considering environmental data not only promise improved support of tourists while traveling in cities, but also help cities to become smarter from many perspectives. Recommendations can be made to avoid currently polluted areas until the air quality improves. The congestions of routes and means of transport can be integrated as route attractiveness attributes to support tourists, locals, and commuters, helping them avoid congested areas. Thus, better distribution of travelers and, in turn, a higher satisfaction of all players in a city can be ensured. Furthermore, cities can analyze how people move between POIs and start initiatives to make alternative routes more attractive by adopting measures such as planting more trees and improving pavements along unpopular routes.

As the next step of our research, we will integrate further route attractiveness attributes into our application. Our user study was limited to a small test area with predefined POIs. For a better understanding of the influence of the attractiveness of routes and detours on a traveler's satisfaction with a recommended trip, we want to evaluate our approach in a larger user study with real tourists and a fully working RS to specify own travel goals and preferred POI categories and to adjust the importance of attractiveness attributes.

REFERENCES

[1] Joan Borràs, Antonio Moreno, and Aida Valls. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications* 41, 16 (2014), 7370 – 7389. https://doi.org/10.1016/j.eswa.2014.06.007

[2] Hieronymus C. Borst, Sanne I. de Vries, Jamie M.A. Graham, Jef E.F. van Dongen, Ingrid Bakker, and Henk M.E. Miedema. 2009. Influence of environmental street characteristics on walking route choice of elderly people. *Journal of Environmental Psychology* 29, 4 (2009), 477 – 484. https://doi.org/10.1016/j.jenvp.2009.08.002

[3] Sef Elshout, Karine Leger, H. Heich, N. Hodges, and F. Nussio. 2007. Www.airqualitynow.eu, a common website and air quality indices to compare cities across Europe. 157–168. https://doi.org/10.2495/AIR070161

[4] Friends of the Earth. 2017. Unmasked: The true story of the air you're breathing.

[5] Damianos Gavalas, Vlasios Kasapakis, Charalampos Konstantopoulos, Grammati Pantziou, and Nikolaos Vathis. 2016. Scenic route planning for tourists. *Personal and Ubiquitous Computing* (2016), 1–19. https://doi.org/10.1007/s00779-016-0971-3

[6] Damianos Gavalas, Michael Kenteris, Charalampos Konstantopoulos, and Grammati Pantziou. 2012. Web application for recommending personalised mobile tourist routes. *Software, IET* 6, 4 (2012), 313–322.

[7] Billie Giles-Corti and Robert J. Donovan. 2003. Relative Influences of Individual, Social Environmental, and Physical Environmental Correlates of Walking. *American Journal of Public Health* 93, 9 (2003), 1583–1589. https://doi.org/10.2105/AJPH.93.9.1583 arXiv:https://doi.org/10.2105/AJPH.93.9.1583 PMID: 12948984.

[8] W. Hawthorn. 1989. Why Ontarians Walk, Why Ontarians Don't Walk More: A Study Into the Walking Habits of Ontarians. Energy Probe, Toronto,.

[9] Daniel Herzog, Christopher Laß, and Wolfgang Wörndl. 2018. TourRec: A Tourist Trip Recommender System for Individuals and Groups. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, New York, NY, USA, 496–497. https://doi.org/10.1145/3240323.3241612

[10] Prashant Kumar, Lidia Morawska, Claudio Martani, George Biskos, Marina Neophytou, Silvana Di Sabatino, Margaret Bell, Leslie Norford, and Rex Britter. 2015. The rise of low-cost sensing for managing air pollution in cities. *Environment International* 75 (2015), 199 – 205. https://doi.org/10.1016/j.envint.2014.11.019

[11] Neville Owen, Nancy Humpel, Eva Leslie, Adrian Bauman, and James F Sallis. 2004. Understanding environmental influences on walking: Review and research agenda. *American Journal of Preventive Medicine* 27, 1 (2004), 67 – 76. https://doi.org/10.1016/j.amepre.2004.03.006

[12] Terri Pikora, Billie Giles-Corti, Fiona Bull, Konrad Jamrozik, and Rob Donovan. 2003. Developing a framework for assessment of the environmental determinants of walking and cycling. *Social Science & Medicine* 56, 8 (2003), 1693 – 1703. https://doi.org/10.1016/S0277-9536(02)00163-6

[13] Hans Pretzsch, Peter Biber, Enno Uhl, Jens Dahlhausen, Thomas RÃűtzer, Juan Caldentey, Takayoshi Koike, Tran van Con, AurÃĺlia Chavanne, Thomas Seifert, Ben du Toit, Craig Farnden, and Stephan Pauleit. 2015. Crown size and growing space requirement of common tree species in urban centres, parks, and forests. *Urban Forestry & Urban Greening* 14, 3 (2015), 466 – 479. https://doi.org/10.1016/j.ufug.2015.04.006

[14] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. 2014. The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT '14)*. ACM, New York, NY, USA, 116–125. https://doi.org/10.1145/2631775.2631799

[15] Brian E. Saelens, James F. Sallis, and Lawrence D. Frank. 2003. Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine* 25, 2 (01 Apr 2003), 80–91. https://doi.org/10.1207/S15324796ABM2502_03

[16] Sherjeel Sikander. 2018. *Integrating Route Attractiveness AttributesInto a Tourist Trip Recommender System.* Master's thesis. Technical University of Munich.

[17] Wouter Souffriau, Pieter Vansteenwegen, Greet Vanden Berghe, and Dirk Van Oudheusden. 2011. The planning of cycle trips in the province of East Flanders. *Omega* 39, 2 (2011), 209 – 213. https://doi.org/10.1016/j.omega.2010.05.001

[18] Takemi Sugiyama, Maike Neuhaus, Rachel Cole, Billie Giles-Corti, and Neville Owen. 2012. Destination and route attributes associated with adults' walking: a review. *Medicine and science in sports and exercise* 44, 7 (2012), 1275–1286.

[19] Theodore Tsiligirides. 1984. Heuristic methods applied to orienteering. *Journal of the Operational Research Society* (1984), 797–809.

[20] Pieter Vansteenwegen, Wouter Souffriau, Greet Vanden Berghe, and Dirk Van Oudheusden. 2011. The City Trip Planner. *Expert Syst. Appl.* 38, 6 (June 2011), 6540–6546. https://doi.org/10.1016/j.eswa.2010.11.085

[21] Pieter Vansteenwegen, Wouter Souffriau, and Dirk Van Oudheusden. 2011. The orienteering problem: A survey. *European Journal of Operational Research* 209, 1 (2011), 1–10.

[22] Pieter Vansteenwegen and Dirk Van Oudheusden. 2007. The mobile tourist guide: an OR opportunity. *OR Insight* 20, 3 (2007), 21–27.

[23] Cédric Verbeeck, Pieter Vansteenwegen, and El-Houssaine Aghezzaf. 2014. An extension of the arc orienteering problem and its application to cycle trip planning. *Transportation Research Part E: Logistics and Transportation Review* 68 (2014), 74–78. http://dx.doi.org/10.1016/j.tre.2014.05.006

[24] Wolfgang Wörndl, Alexander Hefele, and Daniel Herzog. 2017. Recommending a sequence of interesting places for tourist trips. *Information Technology & Tourism* 17, 1 (2017), 31–54. https://doi.org/10.1007/s40558-017-0076-5

[25] M. Xie, L. V. S. Lakshmanan, and P. T. Wood. 2011. CompRec-Trip: A composite recommendation system for travel planning. In *2011 IEEE 27th International Conference on Data Engineering*. 1352–1355. https://doi.org/10.1109/ICDE.2011.5767954

Hybrid Human Machine workflows for mobility management

Eddy Maddalena
University of Southampton
Southampton, United Kingdom
e.maddalena@soton.ac.uk

Luis-Daniel Ibáñez
University of Southampton
Southampton, United Kingdom
l.d.ibanez@soton.ac.uk

Richard Gomer
University of Southampton
Southampton, United Kingdom
r.c.gomer@soton.ac.uk

Mattia Zeni
DISI - University of Trento
Trento, Italy
mattia.zeni.1@unitn.it

Donglei Song
DISI - University of Trento
Trento, Italy
CCST - Jilin University
Changchun, China
donglei.song@unitn.it

Fausto Giunchiglia
DISI - University of Trento
Trento, Italy
CCST - Jilin University
Changchun, China
fausto.giunchiglia@unitn.it

Elena Simperl
University of Southampton
Southampton, United Kingdom
e.simperl@soton.ac.uk

ABSTRACT

Sustainable mobility is one of the main goals of both European and United Nations plans for 2030. The concept of Smart Cities has arisen as a way to achieve this goal by leveraging IoT interconnected devices to collect and analyse large quantities of data. However, several works have pointed out the importance of including the human factor, and in particular, citizens, to make sense of the collected data and ensure their engagement along the data value chain. This paper presents the design and implementation of two end-to-end hybrid human-machine workflows for solving two mobility problems: modal split estimation, and mapping mobility infrastructure. For modal split, we combine the use of i-Log, an app to collect data and interact with citizens, with reinforcement learning classifiers to continuously improve the accuracy of the classification, aiming at reducing the required interactions from citizens. For mobility infrastructure, we developed a system that uses remote crowdworkers to explore the city looking for Points of Interest, that is more scalable than sending agents on the field. Crowdsourced maps are then fused with existing maps (if available) to create a final map that then is validated on the field by citizens engaged through the i-Log app.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Ubiquitous and mobile computing systems and tools**; • **Applied computing** → **Transportation**.

KEYWORDS

modal split, map generation, hybrid workflows, crowdsourcing

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317056

ACM Reference Format:
Eddy Maddalena, Luis-Daniel Ibáñez, Richard Gomer, Mattia Zeni, Donglei Song, Fausto Giunchiglia, and Elena Simperl. 2019. Hybrid Human Machine workflows for mobility management. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3317056

1 INTRODUCTION

The transport sector accounts for 4% of EU's GDP and 9M jobs [17]. The European Commission has outlined congestion, oil dependency, greenhouse gas emissions, and infrastructure quality as major challenges to overcome towards achieving sustainable mobility. From a global perspective, the UN 2030 development agenda specifies as its 11th Sustainable Development Goal (SDG), the provision of "access to safe, affordable, accessible, and sustainable transport systems for all"[1].

A pathway to achieve these goals taken by many cities has being to become *Smart* [1], i.e., take advantage of the connectivity enabled by 4G infrastructure, the ever-decreasing cost of sensors and IoT devices to monitor environmental variables. However, after collecting such large amounts of data, it is of capital importance to (i) ensure that the it *makes sense* for the services that citizens expect to get from the city; and (ii) that citizens are engaged in the process. After all, in many situations, the required data is theirs (their trips, their transport choices, etc.), and only them can provide updates and or corrections. Furthermore, several works have pointed out the importance of supporting citizens in an active role along the whole data value chain [2, 7].

In this paper, we present the design of two end-to-end hybrid human-machine solutions to real mobility problems that municipalities face: the estimation of modal split, and the generation and maintenance of maps of urban mobility infrastructure. The solutions were developed in the context of the EU H2020 project

[1] http://www.slocat.net/sdg-targets

QROWD[13] based on requirements elicited by the Trento Municipality (A partner of the consortium). Modal split is an indicator of the percentage of people that chooses which transportation mode. It is particularly important since it gives a clear picture of how citizens move and commute, and is the first step towards the informed development of transport policies. It is traditionally computed through paper of telephone surveys, in an expensive, time consuming, and non-scalable way. On the other hand, mobility infrastructure maps are important for powering sustainable mobility services, e.g., knowing where bike racks are and their capacity, enable routing services to propose the use of bike, encouraging the reduction of greenhouse emissions. However, for different reasons (infrastructure not owned by the council, digitalisation errors), it is often the case that municipalities don't have that information, and sending municipality employees to scout all the city is expensive.

The remainder of the paper is structured as follows. Section 2 describes i-Log, the mobile application that we use to both collect data from citizens and to interact with them in both hybrid workflows. Section 3 describes the design of the hybrid workflow for mapping mobility infrastructure, and the components that comprise it. Section 4 describes the design of the hybrid workflow for modal split estimation, and the components that comprise it. Finally, section 5 concludes the paper.

2 THE I-LOG MOBILE APPLICATION

How mobile devices can empower users is a relevant topic in areas such as crowdsensing and crowdsourcing [4, 18]. Their unique features make smartphones and wearables the ideal tool for crowdsourcing experiments that require to tap into the wisdom of the crowd and exploiting their knowledge. The general idea is to leverage the power of those devices that are already owned by the citizens, while at the same time balancing the level of intrusiveness of the solutions, to ensure a high rate of response and not hurting the relationship between citizens and public administrations.

With this in mind, the University of Trento developed a mobile application called i-Log [19], which collects data from the user in an unobtrusive, GDPR compliant and efficient way. The application can be used to generate two very diverse types of data, namely (1) streams of value-pairs generated by the devices's internal sensors, while (2) it can also collect the user input in different formats, from text to visual. In crowdsourcing this human knowledge is often provided as annotations or labeling of data collected via sensors [8]. This type of human contribution is at the heart of the "human in the loop paradigm"[2] which leverages both human and machine intelligence to create hybrid machine learning models by involving humans in training, tuning and testing data for a particular machine learning algorithm. The final goal is to use humans to improve the quality of the results of the machine. Specifically, the idea is that every user should provide her own annotations and contributions, that are aware of the context and thus are more meaningful and accurate [3, 6].

A simplified version of i-Log's architecture is presented in Figure 1. The system is composed of a set of modular, logically isolated components, each one enabling a sub-set of the overall functionalities of the application. The modularity of the architecture allows to

[2]https://www.figure-eight.com/resources/human-in-the-loop/

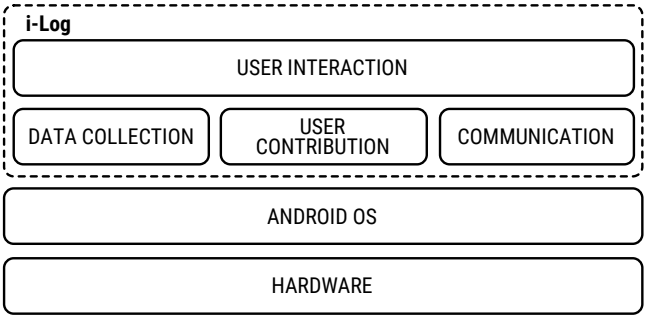


Figure 1: i-Log mobile application architecture.

personalize the application and adapt it to different contexts and projects, with the need to modify only the involved components. This architecture gives i-Log a significant advantage in terms of adaptability and extendibility of its features. The four main components are:

- **Data collection module:** it is responsible for efficiently collecting and storing the data from the smartphone's internal sensors. The data collection has been designed to be remotely configurable in terms of (i) which sensors to use and (ii) at which frequency to collect data from them. In fact, due to the requirements of the individual projects we would need to enable some sensors while disabling others. This is true also from a GDPR perspective where the data minimization principle applies. Once collected, the data are temporarily stored in compressed and encrypted logs file on the device and synchronized over Wi-Fi whenever a connection is available.
- **User contribution module:** is responsible for collecting the user's knowledge in terms of answers to simple questions (a contribution). The knowledge can be of different types, from text, to images and to other objects that are use-case dependent, i.e., coordinates on a map. The questions are sent by a remote server as JSON objects that are then visualized on the smartphone and made available to the user.
- **Communication module:** is responsible for all the outbound and inbound connections. More in details, it allows to contact the backend infrastructure of the application to perform operations such as registering/logging in users, to synchronize the generated logs of data and save them in a database. At the same time, it allows to receive the questions that the user has to reply to provide her own knowledge and keep her in the loop.
- **User interface module:** i-Log's main functionality is to collect data about the user while running in background on the phone. The reason for this is that the collection process must be as less obtrusive as possible. For this reason the user interface is very limited: it consists on a notification system that is always present in the notification area of the smartphone while the data collection is active. This is a mandatory requirement from a GDPR point of view since the user must always been informed when someone is dealing with her data. A second notification is present whenever the user is

asked to provide his knowledge. From these two notifications the user can access the actual views of the application, two menus, Settings and Contributions that allow respectively to setup the application and to have access to all the contributions.

With respect to the QROWD project, i-Log is a pivotal tool to obtain data from citizens of the Municipality of Trento about real time information on traffic and multimodal transport and involve them to improve its mobility by enabling data-driven policy making by providing the users with an interface for harvesting crowdsourced data.

3 GENERATING MOBILITY INFRASTRUCTURE MAPS

An essential ingredient for the design and implementation of mobility policies is knowledge of the current infrastructure. Unfortunately, several public authorities lack such information, since it is expensive to generate, maintain and update, meaning that the location of strategic city items, such as bike racks or disabled parking spots, is unknown. In some cities, engaged communities of volunteers produce and share locations of Points of Interest (an approach studied in the Volunteered geographic information (VGI) field [16]), e.g., contributing to Open Street Maps. Unfortunately, since the participation in such communities is mainly voluntary, there is no control on what their contributions are, when they make them, and on which area of the city they will contribute. An alternative is to physically send municipality employees to locate the required items, however, this does not scale in the area of the city, as to be certain that all the city has been covered, and no items have been over-sighted, several employees need to be sent to the field.

Our solution to this problem combines the automated fusion of existing, possibly incomplete maps (if available), with two crowdsourcing tasks, as described in Figure 2. Our workflow is comprised of four phases: (1) crowdsourced **generation** of new maps (2) **acquisition** of pre-existing and new maps into a central repository, (3) automated **fusion** of the results of available maps to produce an interlinked map and finally, (4) **curation** of the results, using different crowdsourcing strategies.

3.1 Generation

To crowdsource map generation we designed a standalone tool called VCE (Virtual City Explorer). The VCE allows contributors to explore inside street-level imagery services (we used Google Street View[3] in our implementation), and report the coordinates of items that satisfy given requirements. Such contributors can be either citizen volunteers or paid crowdworkers recruited from external channels, e.g., from the traditional crowdsourcing platforms, such as Figure Eight[4] or Amazon Mechanical Turk[5].

The VCE accepts several parameters to control how the mapping is done. An end user, e.g. a municipality, defines the area that needs to be explored by the contributors and the type of infrastructure that wants contributors to locate. In the case of crowdworker contributors, an end user can set how many of them she would like

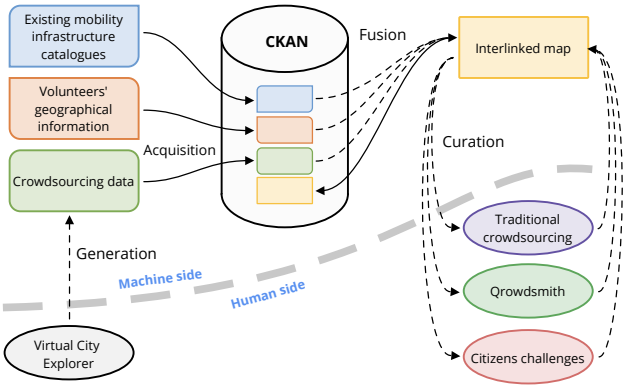


Figure 2: Hybrid workflow for mobility infrastructure mapping.

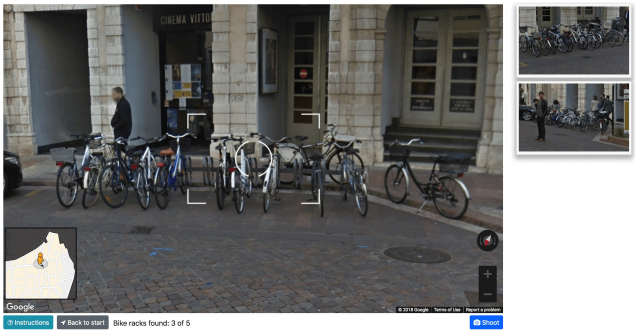


Figure 3: Interface of the VCE from the contributor perspective.

to assign to a given area. Otherwise, for volunteers, the link to the mapping task can be distributed to an arbitrary number of contributors. In this way, the VCE helps to overcome the main limitations of physical approaches, namely, the number of contributors and their direction to specific areas. In turn, the approach is limited to the existence and up-to-dateness of street-level imagery.

Figure 3 shows a screenshot the VCE interface from a contributor's perspective. Before exploration starts, the contributor reads the task instructions that explain its general functioning works and which are the types of objects required to locate. The contributor then starts her exploration from a random within the area of interest. When a contributor discovers a candidate item, she is required to take three photos of it, from three different angles. In the background, the VCE triangulates the vectors of the different angles to determinate the coordinates of the item and stores them in a database. After submitting a pre-established number of items, the task ends. In case a crowdworker was the one completing the task, is redirected to the crowdsourcing recruiting platform to receive her payment.

We evaluated the VCE for the task of generating bike rack maps in comparable areas in the cities of Trento and Washington, finding that contributors are effective locating mobility points of interest

[3]https://www.google.com/streetview/
[4]https://www.figure-eight.com/
[5]https://www.mturk.com/

Compare to sending an employee VCE, also resulting faster covering larger areas, and being less expensive. A separate paper [12] provides a thorough description of the VCE and its experimental evaluation.

3.2 Acquisition

The second phase of our hybrid workflow is data acquisition. In this phase, different maps of PoIs, possibly generated from different sources, are integrated to be used. Such datasets are usually built following different approaches, thus the data collected can have different nature, containing data in different formats and describing different properties (e.g., one approach can represent a bike rack as a geo-polygon, and other as a point). Furthermore, datasets could be updated an arbitrary number of times (e.g., a new execution of the VCE), creating the need of version management.

We use a combination of NiFi processes for data ingestion, and the CKAN open-source DMS (data management system)[6] to provide a centralise repository of datasets for the following phases.

3.3 Data Interlinking and Fusion

From the acquired datasets stored in the CKAN repository in the previous phase, the interlinking and fusion phase is run. First, the interlinking component has to infer which of the entities in the considered datasets refers to the same physical item. The result of the interlinking is a set of entity pairs, each associated with a similarity score (confidence value). This operation that could be performed with a naive approach, requiring computing similarity for all pairs of entities (cartesian product) is done in our interlinking component with *LIMES* [14], an interlinking engine that significantly improves performance by pruning comparisons. Using LIMES, we convert the source data to RDF, generate and derives URI for the PoIs collected in the previous step, and from any other available data source, and represent spatial information using GeoSPARQL. Based on the generated RDF, we use LIMES to generate links of *spatial proximity* among the coordinates of the PoIs, and between any other property, e.g., are they labeled as mobility items, are they popular items, etc. The interlinking phase is followed by the fusion one that concerns with obtaining a single coherent representation of an entity from multiple representations. We implement a clustering operation that first chooses representative identifiers based on the data source ranking and then merge the attributes needed for domain-specific conflict resolution. In the particular case of the city of Trento, we choose as representative the PoIs from data collected by a Muncipality employee, and tried to link to PoIs detected by citizens with i-Log. The result of the interlinking and fusion phases is a new dataset of static items generated through a smart aggregation of the previously available datasets. To verify the results, we use crodsourcing curation, as detailed in Sec. 3.4

3.4 Data curation

The output of the fusion phase is an unified dataset that established which of the mobility infrastructure items collected by different approaches are the same. The interlinking comes with a confidence value to estimate how sure it is that two items are the same. For items with low confidence values, it would be worth to require additional verification. We cal this phase data curation, and developed three human approaches to perform it: (1) by engaging citizens to go verify in place, and (2) recruiting crowdworkers or volunteers for a crowdsourcing task, (3) through a gamification approach that may be played either by crowdworkers or citizens, that can be seen as an extension to (2). The three approaches are detailed in the following three sections.

3.4.1 Citizen challenges. The first approach we can use as a confirmation of the interlinking output is to leverage directly on the citizens. A difference with respect to the other solutions is that the citizens need to physically go and check if a bike rack is still present or not. This can be considered an additional challenge with respect to a solution where a person can validate from her computer, but at the same time gives us additional confirmation, in real time and with the most up-to-date information. This type of data curation is made available through the i-Log application and its functionality that allows to present questions to the users. More in detail, the citizens can participate and contribute to *Challenges* made available by the Municipality of Trento (Fig. 4 (left)). Once they agreed to participate, they are presented with a map that highlights where the bike racks are located (from the interlinking step).

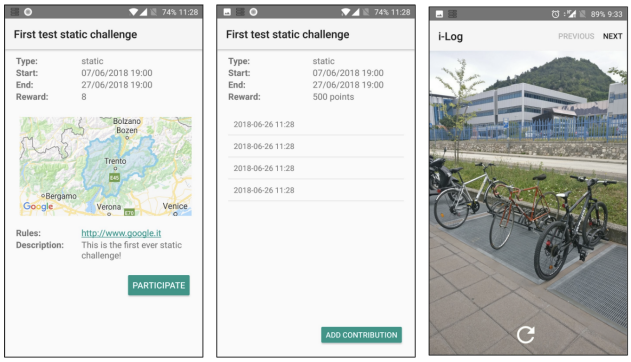


Figure 4: Three i-Log interfaces, for (left) a user to decide if accept to participate a challenge, (middle) a user contribute with a new item detection, and (right) a user taking a picture of a new item discovered.

The challenge consists in physically going to the designated points and confirm if the bike racks are really there. The confirmation occurs through what we call a *contribution* (Fig. 4 (center)) that consists in the following actions that the user has to take:

(1) Read the textual/visual instructions that are provided through the i-Log user interface. The message we were presenting is: *Hi! It seems that there is a bike rack here; is it true? If yes, before answering the questions on the bike rack, we would like to ask you if you could localize yourself, then take three nice picture of the bike rack itself. Please note that the photo should cover the whole bike rack!.*
(2) Localize her device through a dedicated procedure. It consists in a map view that gets the location from the GPS sensor and generates a pinpoint. The location is considered valid only if the attached accuracy is below $10.0m$.

[6] https://github.com/ckan/ckan

(3) Reply to a first textual question: *Is the bike rack still here?*. The available answers are Yes/No. If the selected answer is No, then the contribution is considered finished.
(4) Take a picture of the bike rack. The location from the GPS sensor is also collected at this point 4 (right)).
(5) Take a second picture in a position different from the first one.
(6) Take a third picture in a position different from the first two.
(7) Reply to a textual question: *What kind of bike rack do you see?*. In this case the answers are three pictures of three different types of bike racks.
(8) Reply to a textual question: *How many spots does the bike rack have?*.
(9) Reply to a textual question: *How many available spots does the bike rack have?*.

Once a contribution has been generated, i-Log stores them locally before sending them to the backend once a Wi-Fi connection is available.

3.4.2 Traditional crowdsourcing. The second approach we can use to confirm the output of the interlinking component is through a simple crowdsourcing task to be carried out either by paid crowdworkers or volunteer contributors.

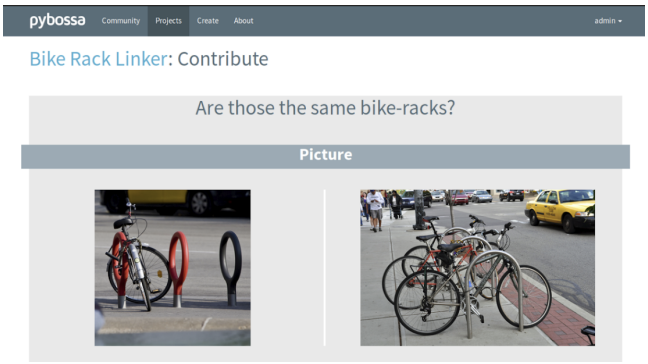


Figure 5: The data curation phase requires crowdworkers to carry out simple tasks, such as determining if the interlinking of two items from different dataset was performed correctly.

Fig. 5 shows, the interface of a task we designed to be run in Pybossa[7], a crowdsourcing framework to analyze or enrich data. The task requires contributors to confirm if the items pictured in two photos (bike racks in Fig. 5) correspond to the same item. .

3.4.3 Qrowdsmith. An extension to the previous approach is the use of gamification to better engage with contributors. The effort of creating a gamified environment for crowdsourcing contributors is justified by studies that demonstrate how hedonic motivation can be placed alongside the economic one to lead to better quality results, by improving the quality of the engagement of the contributors [5]. In this context, we developed Qrowdsmith, a standalone tool that makes use of gamification for engaging with contributors, either volunteers or paid crowdworkers recruited from traditional crowdsourcing platforms such as figure eight or Amazon Mechanical Turk. Qrowdsmith offers a series of gamified elements such as leaderboards, badges and levels to engage with crowdworkers providing them with feedback concerning the work done over time to favour the establishment of long-time relations with workers, that are even more motivated to return on the platform to increase their scores and climb the leader board. Fig 6 shows the interface of the main menu of Qrowdsmith: the crowdworker can choose among several available tasks, some of them to be carried out in the solo mode, other requiring participation of multiple users. Currently, Qrowdsmith is in a prototypal stage and supports three types of tasks, respectively: (i) *image tagging* where contributors are required to provide labels to describe the content of given pictures; (ii) *item comparision* that similarly to the task described for the Traditional crowdsourcing apprach (Sec. 3.4.2), requires workers to establish if two photos refer to the same item; and (iii) *item validation* where contributors are required to judge whether an item depicted on a photo conforms to the given guideline.

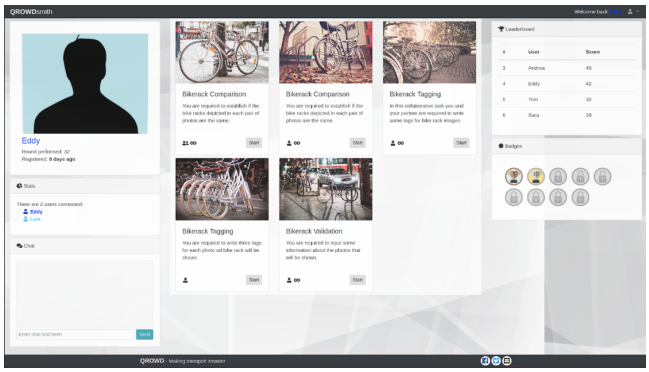


Figure 6: The interface of Qrowdsmith, where volunteers or paid contributors engage performing tasks in an environment having gamified elements, such as badges, leaderboards, and rewards.

4 MODAL SPLIT ESTIMATION

In this section we describe the hybrid workflow for estimating modal split[8] Figure 7 depicts the general schema of the solution. We follow approximately the same general pattern as described in [15]. The general idea is to use i-Log to collect raw sensor data about citizens movement, and compute as automatically as possible the trips and transport modes of the citizen. Citizens might be required to confirm the computed trips and their transport mode if the machine is not confident enough. This turns the lengthy and cumbersome process of filling a paper survey into just letting an app run in the background and occasionally answer some questions

The next sections details the main phases of the hybrid workflow for modal split: the personal data collection (Sec. 4.1), the

[7] https://pybossa.com/

[8] A demonstration of an earlier version of this workflow was presented in the Project Showcase Track of KDD'18. https://www.kdd.org/kdd2018/files/project-showcase/KDD18_paper_1810.pdf

trip inference 4.2, and finally the trip and segment confirmation (Sec. 4.3).

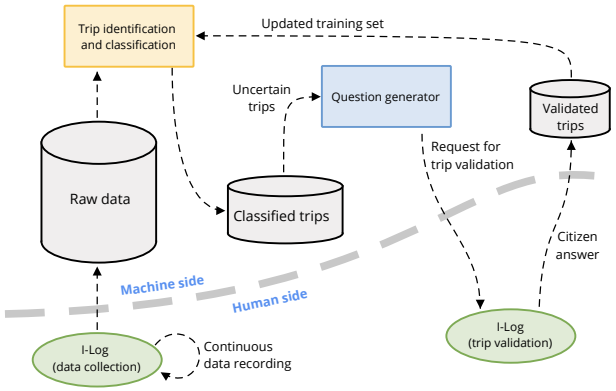


Figure 7: General schema of the workflow for the computation of the personal modal split.

4.1 Data collection from citizen devices

As explained in Section 2, the i-Log application can be used to collect personal big data from the mobile devices of the users. The fact that no dedicated device has to be used and, instead, the citizens can use their own device facilitates the exploitation of the tool in real life and in the wild, scenarios like the ones developed in the QROWD project. i-Log is a general purpose data collection tool that adapts to the specific need of the use case of each project. This is true for both the type of data it can collect, sensor data and user's contributions. The adaptation involves multiple dimensions: (1) the analysis of the privacy and ethical requirements, (2) the definition of the devices to collect data from, (3) the definition of the sensors on each device to collect data from and (4) the definition of how frequently to collect data from each sensor.

When dealing with people, it is of primary importance to take into consideration privacy and ethics. Each institution usually has a dedicated board that deals with such topics when humans are involved, and this happened also for the QROWD project. The outcome of the discussion is an installation procedure that is intended to present the user the *Informed Consent* about the data collection and explicitly ask her to *grant permissions* to collect data from those, more critical sensors from a privacy perspective, i.e., the GPS. The procedure guides the user through the process with images and text that explains the reasons for which the study needs the specific kind of data.

Another crucial aspect when dealing with a study in the wild relates to the fact that is hard to foresee which devices the citizens will use to participate. With respect to a closed environment study in a laboratory setting, this element creates uncertainly and can lead to additional problems and issues that are device specific. Usually in this kind of studies the participants are given a single specific device that has been extensively tested, while in our study this was not possible. i-Log tries to overcome this problems because has been designed to work on any Android device with some very limited exceptions.

The focus of the Modal Split use case within the QROWD project relates to understanding how people move around the city. Starting from this we defined the requirements in terms of the data to be collected, which mainly belongs to the inertial sensors. We finally decided to leverage on the *GPS*, *accelerometer* and *gyroscope* sensors. We added an additional virtual sensor that could help in detecting the *activity performed* by the user that is provided on the device as s Google service. This sensor returns a label among "running", "walking", "still", "on a bike" and "inVehicle" with an attached accuracy value, in percentage, from 1 to 100. The selection of only these few sensors with respect to all the ones available in i-Log perfectly matches the principle of data minimization required by GDPR, that is defined as *"Personal data shall be... (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"*.

The other element that has to be specified when using i-Log as data collection tool is the frequency at which the data are generated by the internal sensors. Similarly to the previous point, this is highly dependent on the use case but can also be capped by external conditions, i.e., limited storage size, limited internet connectivity, battery constraints, among others. For the Modal Split within the QROWD project we didn't have particular constraints and we decided then to go for the most efficient solution, that balances the amount of data collected and the recognition performances, being $20Hz$ for the inertial sensors [9]. The motivation for this relies in the Shannon–Nyquist theorem as reported by [10], that states that for a successful loss-less reconstruction of a particular signal, the data needs to be sampled with at least twice its highest frequency, being $10Hz$ for human movements such as walking, running, or cycling. Concerning instead the other sensors, namely the GPS and the activity performed, the frequency has been set to one value generated every minute or $1/60Hz$. The reason why this frequency is much lower with respect to the inertial sensors differs from one to the other. For the GPS the main limitation was the battery capacity of the devices, in fact, this sensor is the most energy demanding one of modern smartphones [11]. In order to don't affect the user experience by draining the battery of the device, we decided to sacrifice some location updates and obtained a full day battery life. On the other hand, for the activity performed, the limit was imposed by Google that to not allow to have higher updates.

Table 1: Sensors and frequency recorded by i-Log

Sensor	Accelerometer	Gyroscope	GPS	Activity
Frequency	$20Hz$	$20Hz$	$1/60Hz$	$1/60Hz$

Table 1 summarizes the sensors use for the Modal Split case study with the corresponding logging frequency.

The citizens were required to use the i-Log application on their mobile devices for a period of two weeks, every day for as many hours as possible in order to detect all their trips in the city. At any moment, the user could stop the data collection in compliance with GDPR and restart it whenever he feels to.

4.2 From raw sensors data to trips

In an average day, i-Log collects about 100Mb of raw data produced by the sensor installed on the devices of its users. i-Log sends daily such data to the database for the storage. At this stage, such data are still raw; thus difficulty interpretable. For some data streams, trip identification and modal classification can be quite easy, e.g., a series of subsequent points, detected over a restricted amount of time, having coordinates that perfectly overlap those of a railway, can be inferred as a train trip. Nevertheless in some other cases, such inference is uncertain, or even impossible. Data collected by i-Log might be inaccurate or incomplete due to the technical limitations of smartphones described in section 4.1, representing a challenge for the learning component. Another issue that needs to be taken in consideration since highly relevant to the modal split computation is the presence of multi-modal trips. Multi-modal trips are trips in which the citizen changed the transportation mode multiple time from the starting point or the trip to the destination one. We call *segmentation* to the action of splitting a multi-modal trip into more single-modal trips. Correct segmentation is challenging, as transport mode changes tend to happen in a short period of time, making difficult to identify the exact point in time and space of the change.

Several works have looked at the implementation of Machine Learning classifiers to infer, for each day of data, the trips that citizens took and what was the transport mean used. Such classifiers can be studied in isolation, or in the context of a general workflow where results of the classification are sent to citizens in order to validate them, and use the citizen's input to implement a reinforcement learning scheme. We follow the latter scheme, as described in our previous work [13].

As purely machine classifiers, we implemented Decision Trees and Random Forest on top of gyroscope and accelerometer data. One of the challenges faced is to generate enough training data on the particular topology and available transport means of the target city. Our answer to that challenge is to start training on publicly available data, like the UCI Human Activity Recognition dataset or the, and then, using a reinforcement pipeline, increase the accuracy of the model progressively, each time a travel survey is applied. During the first iterations, the accuracy is low, and we expect to have to ask many questions to citizens to confirm and update their trips. The classifier is retrained after each survey (2 weeks worth of data collection on 20 to 600 citizens), progressively increasing its accuracy.

Our approach makes very important the correct management of the interface with citizens, as they are the only ones that can provide the real answer of the trip. As such, we analysed the most common errors committed by the Machine Learning classifier, in order to provide an appropriate interface for each of them. Figure 8 shows the five possible outputs of the trip segmentation process: (i) *No errors*, the segmentation is correct and perfectly in line with the trip made by the citizen; (ii) *wrong transport change point*, the transport modes are detected correctly, but the change point is wrong; (iii) *wrong transport mode*, the multi-modal trip is correctly segmented, but one or more transportation modes inferred are wrong; (iv) Under segmentation, the multi-modal trip in segmented into less segment than the expected (two or more segments are merged together); and (v) *over segmentation*, the multi-modal trip in segmented into more segment than the expected (some segment are wrongly divided instead on being together).

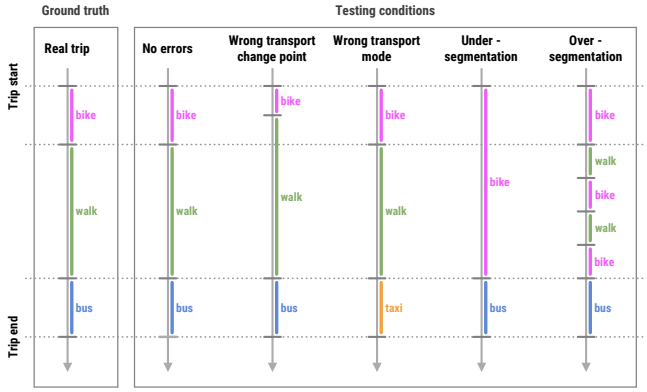


Figure 8: The segmentation process can lead to five possible outputs.

4.3 Trip confirmation

The capability of the classifier in generating an effective learning model, that allows identifying trips correctly, depends on the quality of the training set. In particular, an effecting tuning of machine learning classifiers requires a training set enriched with correct classifications. Initially, the classifier works with data collected from other cities, however, to enable continuous improvement, we implemented a feature that allows requesting for a trip classification directly to the citizen who generated that trip (usually in the previous 24 hours), via the i-Log app. The feature also allows municipalities to collect quality data for the modal split estimation from the first run.

i-Log allows to collect human feedback in terms of answers to specific questions administered through the application. These questions can be simple textual questions that require the user to select one among different multiple-choice answers, but can require the user to take more complex actions, such as localize herself, taking a picture or interact with maps. The questions are triggered by a *REST API* call to a backend server that leverages on Google Firebase [9] to deliver them to the devices. The principle behind the service is that Google opens a connection with the device when it is convenient to do so and in this window it delivers messages sent by multiple applications simultaneously. The messages for confirming modal split trips are composed by a JSON string that contains text and metadata. Similarly to the contributions for the challenges described in Section 3.4.1, the answers are temporarily stored on the device and synchronized with the backend server whenever a Wi-Fi connection is available. Figure 9 shows i-Log interface when requiring to a citizen to confirm a trip. On the left the view that asks about what the citizen did in the selected location. In the center, the citizen is asked to select which transportation mean she used for the highlighted trip. Finally, on the right, the

[9] https://firebase.google.com/

user is required to put a marker on the location where he changed his transportation mode.

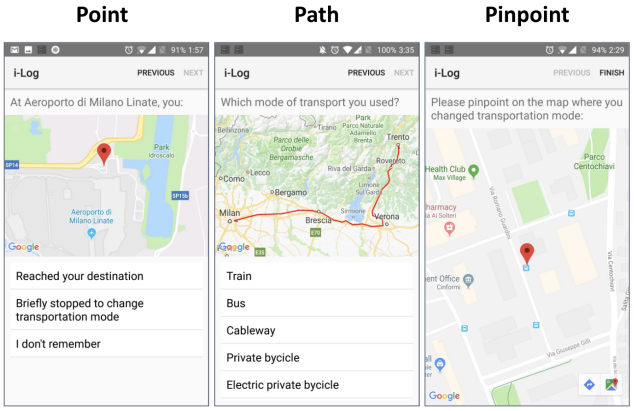


Figure 9: The interface of i-Log when requiring to a citizen to confirm a trip.

5 CONCLUSIONS

In this paper, we presented two human-machines hybrid workflows that make use of human computation strategies to engage with crowds (both citizen with a local knowledge of the territory and remote crowdworkers) to tackle two mobility related issues: the problem for public authorities of completing the mobility infrastructure maps and the estimation of the modal split. The solutions presented are part of a big picture in the context of the ongoing QROWD European Project. The studies discussed varies in several factors, such as the type of crowd, engagement required, and costs to sustain. Our ongoing research covers such aspects, aiming at developing a homogeneous integration of the proposed services in order to provide a platform for end users, who might need to use crowdsourcing services even without having specific technical knowledge. On one side, we used the i-Log application to engage directly with the citizens, and leverage on their data and knowledge to achieve the defined goals. On the other side, we leveraged traditional crowds.

Our preliminary results on the mobility infrastructure workflow [12] show that the crowdworkers can complement current datasets very well. However, this is a limitation for areas that are not currently covered by an on-field agent. Further research is required to identify areas where remote crowdworkers cannot reach, and combine with appropriate spatial crowdsourcing techniques to send field agents only to those locations.

With respect to the Modal Split workflow, our initial experiments show that our machine learning classifiers show similar accuracy than state of the art ones, however, better strategies to drive the reinforcement learning need to be developed, in order to boost the effect of making many surveys. Interface-wise, our approach was very effective for confirming true trips and fixing positional errors, but was not so effective to solve oversegmentation types of errors (especially when there are many segments). Further research is required to design an appropriate UI for this case.

ACKNOWLEDGMENTS

This paper is supported by the QROWD project, part of the Horizon 2020 programme under grant agreement 732194. We would like to thank the crowdworkers who participated in this work, and the team at Municipality of Trento for the support provided.

REFERENCES

[1] Vito Albino, Umberto Berardi, and Rosa Maria Dangelico. 2015. Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology* 22, 1 (2015), 3–21. https://doi.org/10.1080/10630732.2014.942092

[2] Tuba Bakıcı, Esteve Almirall, and Jonathan Wareham. 2013. A Smart City Initiative: the Case of Barcelona. *Journal of the Knowledge Economy* 4, 2 (01 Jun 2013), 135–148. https://doi.org/10.1007/s13132-012-0084-9

[3] Paolo Bouquet and Fausto Giunchiglia. 1995. Reasoning about theory adequacy. a new solution to the qualification problem. *Fundamenta Informaticae* 23, 2, 3, 4 (1995), 247–262.

[4] Vincenzo Della Mea, Eddy Maddalena, and Stefano Mizzaro. 2013. Crowdsourcing to Mobile Users: A Study of the Role of Platforms and Tasks.. In *DBCrowd*. Citeseer, 14–19.

[5] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 2015. Improving Paid Microtasks Through Gamification and Adaptive Furtherance Incentives. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 333–343. https://doi.org/10.1145/2736277.2741639

[6] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. 2017. Personal context modelling and annotation. In *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 117–122.

[7] Daniel Gooch, Annika Wolff, Gerd Kortuem, and Rebecca Brown. 2015. Reimagining the Role of Citizens in Smart City Projects. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1587–1594. https://doi.org/10.1145/2800835.2801622

[8] Bin Guo, Zhiwen Yu, Xingshe Zhou, and Daqing Zhang. 2014. From participatory sensing to mobile crowd sensing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE, 593–598.

[9] Holger Junker, Paul Lukowicz, and Gerhard Troster. 2004. Sampling frequency, signal resolution and the accuracy of wearable context recognition systems. In *Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, Vol. 1. IEEE, 176–177.

[10] Aftab Khan, Nils Hammerla, Sebastian Mellor, and Thomas Platz. 2016. Optimising sampling rates for accelerometer-based human activity recognition. *Pattern Recognition Letters* 73 (2016), 33 – 40. https://doi.org/10.1016/j.patrec.2016.01.001

[11] A Tawalbeh Lo'ai, Anas Basalamah, Rashid Mehmood, and Hala Tawalbeh. 2016. Greener and smarter phones for future cities: Characterizing the impact of GPS signal strength on power consumption. *IEEE Access* 4 (2016), 858–868.

[12] Eddy Maddalena, Luis Daniel Ibáñez, and Elena Simperl. 2019. On the mapping of Points of Interest through StreetView imagery and paid crowdsourcing. *CoRR* abs/1901.09264 (2019). arXiv:1901.09264 http://arxiv.org/abs/1901.09264

[13] Eddy Maddalena, Luis-Daniel Ibáñez, Elena Simperl, Mattia Zeni, Enrico Bignotti, Fausto Giunchiglia, Claus Stadler, Patrick Westphal, Luís PF Garcia, and Jens Lehmann. 2018. QROWD: Because Big Data Integration is Humanly Possible. In *Proceedings of the Project Showcase Track of KDD2018*.

[14] Axel-Cyrille Ngonga Ngomo and Sören Auer. 2011. LIMES: A Time-efficient Approach for Large-scale Link Discovery on the Web of Data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three (IJCAI'11)*. AAAI Press, 2312–2317.

[15] Adrian Corneliu Prelipcean, Yusak Octavius Susilo, and Gyözö Gidofalvi. 2018. Future directions of research for automatic travel diary collection. In *DIVA*. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-227256

[16] Daniel Sui, Sarah Elwood, and Michael Goodchild. 2012. *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Springer Science & Business Media.

[17] The European Commission. 2018. *Transport in the European Union*. The European Commission. https://ec.europa.eu/transport/sites/transport/files/2018-transport-in-the-eu-current-trends-and-issues.pdf

[18] Dejun Yang, Guoliang Xue, Xi Fang, and Jian Tang. 2012. Crowdsourcing to Smartphones: Incentive Mechanism Design for Mobile Phone Sensing. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*. ACM, New York, NY, USA, 173–184.

[19] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. 2014. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 299–302.

Intelligent Traffic Congestion Classification System using Artificial Neural Network

Md Ashifuddin Mondal
Department of Computer Science and Engineering
Narula Institute of Technology
Kolkata, India
ashifuddin.mondal@nit.ac.in

Zeenat Rehena
Department of Computer Science and Engineering
Aliah University
Kolkata, India
zeenatrehena@yahoo.co.in

ABSTRACT

Managing the ever increasing road traffic congestion due to enormous vehicular growth is a big concern all over the world. Tremendous air pollution, loss of valuable time and money are the common consequences of traffic congestion in urban areas. IoT based Intelligent Transportation System (ITS) can help in managing the road traffic congestion in an efficient way. Estimation and classification of the traffic congestion state of different road segments is one of the important aspects of intelligent traffic management. Traffic congestion state recognition of different road segments helps the traffic management authority to optimize the traffic regulation of a transportation system. The commuters can also decide their best possible route to the destination based on traffic congestion state of different road segments. This paper aims to estimate and classify the traffic congestion state of different road segments within a city by analyzing the road traffic data captured by in-road stationary sensors. The Artificial Neural Network (ANN) based system is used to classify traffic congestion states. Based on traffic congestion status, ITS will automatically update the traffic regulations like, changing the queue length in traffic signal, suggesting alternate routes. It also helps the government to device policies regarding construction of flyover/alternate route for better traffic management.

CCS CONCEPTS

• **Networks → Network services**; **Cloud computing**; • **Information systems → Information systems applications**; **Decision support systems**; *Expert systems*; *Data analytics*;

KEYWORDS

Intelligent transportation system; traffic congestion classification; artificial neural network; IoT; smart city

ACM Reference Format:
Md Ashifuddin Mondal and Zeenat Rehena. 2019. Intelligent Traffic Congestion Classification System using Artificial Neural Network . In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317053

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317053

1 INTRODUCTION

With the rapid development of information and communication technology (ICT), the concept of "smart city" emerges as a way to enhance urban management and environmental sustainability as well as increase standard of living of citizens [1]. Smart mobility is one of the important aspects in "smart city" [10]. Along with the population burst, enormous vehicular growth, socio-economical growth and rural to urban migration has imposed a very difficult challenge to the traffic management system especially in developing countries like India [15]. Population and economic growth leads to enormous vehicular growth in the streets of urban areas. The number of registered vehicles has increased from 0.3 million in 1951 to about 142 million in 2011 in India [14]. Almost every country throughout the world is facing problems in managing the transportation infrastructure and facilities [17]. So, ever increasing vehicular growth puts lots of pressure on transportation infrastructures whatever the country currently have. The streets of Indian urban areas are not able to accommodate the large number of vehicles and thus lead to traffic congestion. Increase in air pollution, loss of valuable time and money of the citizens are the common consequences of traffic congestion in the roads of urban areas. India experiences monetary loss of $6 billion a year due to traffic congestion according to World Bank study [14]. Also high traffic jam results in the increase in number of road accidents [4]. Hence a smart system is required for identification of traffic congestion state of a road to assists the users. IoT based Intelligent Transportation System can be used for monitoring and managing traffic congestion to achieve smooth mobility of traffic, reducing the travel time and air pollution as well enhance the road safety of commuters [15].

Identifying the traffic congestion status of the different road segment is one of the important tasks for the better management of traffic. Traffic congestion status identification over time series helps the traffic regulation authorities to make proper decisions regarding traffic rules and regulations of transportation system. A system needs to be devised which will monitor the traffic continuously by capturing the different parameters of traffic of a road segment and identifies the congestion status. Here,IoT plays a significant role in modern traffic management [11].

The aim of this paper is to present an intelligent traffic congestion status classification framework based on artificial neural network (ANN) [2]. The proposed system acquires real time traffic information and determines the traffic congestion status of different road segments using ANN technique. It will get the real time traffic information like traffic density and average speed of the vehicles from different road segments with the help of in-road stationary sensors and then infer the traffic congestion status using ANN

technique. The traffic pattern will help the government/traffic regulatory bodies to make a proper decision regarding traffic rules and regulations and improvement in the area of traffic infrastructure.

The rest of the paper is organized as follows. Section 2 highlights related work on traffic data collection techniques and traffic congestion status estimation. In Section 3, the ANN based proposed system design for traffic congestion estimation and classification is described. Section 4 presents the simulation and result analysis of traffic congestion status identification. Finally, we conclude our paper in Section 5.

2 RELATED WORK

Due to the urbanization, the demand for good transportation system increases tremendously. Therefore, the number of vehicles in road also has risen day by day. There are a lot of research works have been done in different applications of smart traffic management system. Identifying traffic congestion on road is one of the important aspects of it.

In [18], the authors proposed a technique for aggregating floating car data (FCD) to reconstruct accurate traffic state. They determine traffic density and traffic flow with the help of traditional stationary devices like video camera, loop detector and FCD. An efficient and large scale traffic monitoring system is presented in [16], which deduce spatio-temporal traffic data by detecting cyclist, vehicles, pedestrian's etc. using radio-based Bluetooth/Wi-Fi technology. The traffic object equipped with Blue-tooth/Wi-Fi enable device, can detect all other traffic objects by their Bluetooth/Wi-Fi identification number. An object/vehicle is observed or detected by multiple other objects in different location and time and sends the information to the processing unit which reconstruct the traffic trajectory and estimates traffic situation.

In [19], authors proposed a technique to classify road traffic congestion level using decision tree algorithm. They consider vehicle velocity as a parameter to identify congestion level. They collect the road traffic data using GPS device. They used sliding window technique to generate moving pattern from vehicles velocity. They have used J48 decision tree algorithm to develop decision tree model to classify the congestion level. In [7], speed performance index was used to evaluate the road network traffic congestion states. Authors consider vehicle's speed as an important parameter to estimate traffic states. The speed performance index is calculated from vehicle's speed and based on index value, system classify the traffic congestion states. Coifman et al. [3] proposed a method which processes the data collected from automatic vehicle location system (AVL) to measure the travel time and average speed over the freeway and in turn it determines the traffic condition. In another study [13], traffic congestion estimation algorithm has been proposed where congestion features are extracted from MPEG video data and then use Gaussian Hidden Markov model to determine traffic congestion level. In [9], fuzzy logic rule has been used to estimate congestion level. Also in [8] fuzzy logic has been used to detect traffic congestion and a fuzzy based model is used for traffic control.

According to the above mentioned studies, some disadvantages are found in determining the traffic congestion states like: lack of congestion computational model, selection of traffic parameters to infer congestion states. In [9], vehicle velocity is only considered as a parameter to classify traffic congestion level. Similarly, in [8], authors only considers speed performance index as a measuring factor of congestion assessment. It also lacks proper computational technique. To overcome these short comings, this paper proposes an intelligent traffic congestion classification system based on ANN. The system considers two traffic parameters: traffic density and average speed to determine congestion states. It uses ANN model as computational technique to classify congestion levels.

3 PROPOSED SYSTEM

Intelligent traffic monitoring system is one of the important aspects of intelligent transportation system (ITS). It is the system which is administrated by government or transportation authority and can be used to monitor as well as analyze the traffic status of different road segments. The system basically collects the traffic data in real time and determines the congestion states for decision making and management of traffic in an efficient way. The proposed system classifies the traffic congestion states of road segments into three categories: High Congestion, Medium Congestion and Free Flow. Here, in-road stationary sensors capture the different traffic parameters like speed of vehicle, number of vehicle in road in real time and send it to the Artificial Neural Network based information processing unit, which determines traffic congestion states of road segments.

3.1 Overview of the System

The proposed system composed of components namely data collection unit, data transmission unit and information processing unit. These are described as follows.

3.1.1 Data Collection Unit. Traffic data is very important for decision making and precise traffic management planning. The proposed system uses in-road stationary sensors to monitor real time traffic. The sensors are deployed on the road at the both end of road segment. The sensors will capture the speed of vehicles, number of vehicles on the road by the stationary sensors.

3.1.2 Data Transmission Unit. The sensed traffic data need to be sent in periodic manner from field to remote data analysis unit. The local data collection unit will send the aggregated data through wireless or wired communication.

3.1.3 Information Processing Unit. The raw traffic data collected from the field are used to calculate average speed of the vehicles and traffic density of the road. These are analyzed to infer knowledge regarding traffic congestion states of different road segments so that regulatory authorities can make proper planning and decisions in timely manner.

The proposed system uses artificial neural network (ANN) in attempt to make the information processing unit intelligent. The unit receives the real time traffic data which are fed into ANN and it determines the traffic congestion status. The Figure 1 shows the overall design of intelligent traffic congestion classification system.

3.2 Traffic Parameters

The two main traffic parameters are used in the proposed system. These are namely i) traffic density and ii) average speed of vehicles

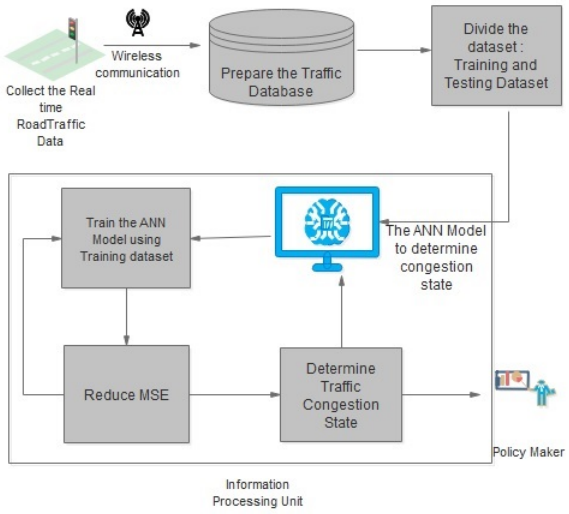


Figure 1: Intelligent Traffic Congestion Classification System.

of each road segment. These are defined as follows.

Traffic Density: Traffic Density is defined as the number of vehicles occupying a given length of a road segment. It can be expressed as:

$$d = \frac{n}{l} \tag{1}$$

where, n is the number of vehicles and l is the length of the road segment.

Average Speed: Average Speed (s_{avg}) is the sum of speed of all the vehicles divided by total number of vehicles.

where, Speed (s_i) is defined as the total distance travelled by the vehicle per unit time. Hence Average Speed can be expressed as:

$$s_{avg} = \frac{\sum_{i=1}^{n} s_i}{n}, \quad where \quad s_i = \frac{dis_i}{t} \tag{2}$$

where, dis_i is the distance travelled by i^{th} vehicle and t is the time period.

3.3 Traffic Congestion States Determination

ANN is an important branch in the field of artificial intelligence [2]. It is a good solution to make an information processing unit to become intelligent. Inspired by biological neural network, ANN based system was developed to solve different complex problems and can be used efficiently to categorize the congestion states. The ANN is composed of large number of elementary processing units called "neurons" and they are connected with each other with some weights. Each "neuron" takes input from source of information and produces output with the help of transfer function. Generally, ANN consists of three layers: i) input layer, ii) hidden layer and iii) output layer and each layer composed of several "neurons". Each layer is interconnected with some weights. In this work, fully connected multilayer perceptron (MLP) [5][6] is used.

Here, the input layer receives two traffic parameters i.e. traffic density and speed of vehicle along with road segment identification number as input to the neural network. This layer only passes the information from real world to the hidden layer, no computation is performed here.

Hidden layer consists of several neurons and it performs the computations on in-coming data and transfer the information to the output layer. An ANN may have any number of hidden layers, but here we consider only one hidden layer having five neurons.

Output layer is responsible for providing output to the outside world. As this work classifies the congestion states into three categories namely High congestion, Medium congestion and Free flow, the proposed ANN model have three neurons each corresponds to different congestion states. Figure 2 shows the proposed ANN model to classify congestion status. It has three inputs $I = I_1, I_2, I_3$ (I_1=Traffic Density, I_2= Average Speed and I_3= Road Segment Identification).

It is clear from the figure that each node of input layer is connected to every node of hidden layer with some weight. Also each node of hidden layer is connected to every node of output layer with some weight. The weighted matrix from input layer to hidden is represented by W_1 and weighted matrix from hidden layer to output layer is represented by W_2. Matrix W_1 and W_2 are shown as follows where $w^i_{j,k}$ represents the weight of the connection to the j^{th} neuron from k^{th} neuron in the i^{th} layer.

$$W^1 = \begin{bmatrix} w^1_{1,1} & w^1_{1,2} & w^1_{1,3} \\ w^1_{2,1} & w^1_{2,2} & w^1_{2,3} \\ w^1_{3,1} & w^1_{3,2} & w^1_{3,3} \\ w^1_{4,1} & w^1_{4,2} & w^1_{4,3} \\ w^1_{5,1} & w^1_{5,2} & w^1_{5,3} \end{bmatrix}$$

$$W^2 = \begin{bmatrix} w^2_{1,1} & w^2_{1,2} & w^2_{1,3} & w^2_{1,4} & w^2_{1,5} \\ w^2_{2,1} & w^2_{2,2} & w^2_{2,3} & w^2_{2,4} & w^2_{2,5} \\ w^2_{3,1} & w^2_{3,2} & w^2_{3,3} & w^2_{3,4} & w^2_{3,5} \end{bmatrix}$$

Every neuron (node) in the hidden layer processes the information based on input and bias using weighted sum method [5] and produces output based on activation function [5].

The weighted sum (n^1_1) at 1^{st} node in the hidden layer represented by :

$$n^1_1 = w^1_{1,1} * I_1 + w^1_{1,2} * I_2 + w^1_{1,3} * I_3 + b \tag{3}$$

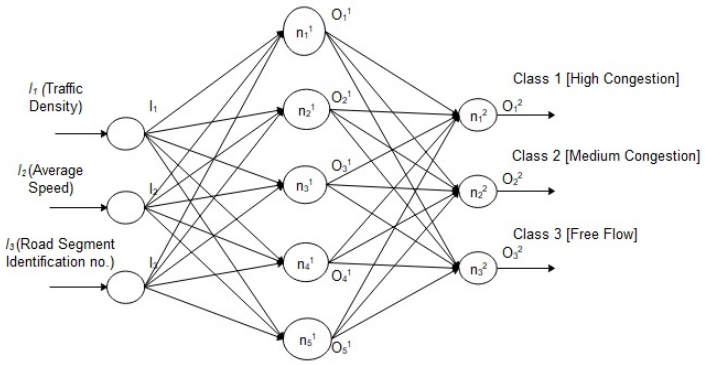


Figure 2: Artificial Neural Network Model.

Then the activation function is applied on the weighted sum (n_1^1) to produce the net output (O_1^1) at 1^{st} node in the hidden layer and can be represented by:

$$O_1^1 = f(n_1^1) \tag{4}$$

where f is activation function. This paper uses log sigmoid activation function [5] which maps the output in the range from 0 to 1 and is represented by :

$$f(n_1^1) = 1/(1 + e^{-n_1^1}) \tag{5}$$

Hence, weighted sum (n_1^1) at 1^{st} node in the hidden layer can be represented in dot product form:

$$n_1^1 = \sum_{k=1}^{m} w_{1,k}^1 i_k + b, \quad where \quad m = 3 \tag{6}$$

The weighted sum (n_1^1) is fed to log sigmoid activation function to get the net output (O_1^1) at the 1^{st} node in the hidden layer which can be represented in dot product form as follows:

$$O_1^1 = f(\sum_{k=1}^{m} w_{1,k}^1 i_k + b) \tag{7}$$

Likewise, net output $O_1^j, j = 5$ for all five node in the hidden layer can be calculated.

Now hidden layer output is fed into the node of output layer to calculate the final output for every node of output layer.

The weighted sum (n_1^2) at 1^{st} node (Class 1 node) in the output layer can be represented in dot product form:

$$n_1^2 = \sum_{k=1}^{m} w_{1,k}^2 O_k^1 + b, \quad where \quad m = 5 \tag{8}$$

Then weighted sum (n_1^2) is fed to log sigmoid activation function to get the net output O_1^2 of 1^{st} node (Class 1 node) in the output layer.

The net output O_1^2 at the 1^{st} node in the output layer can be represented in dot product form:

$$O_1^2 = f(\sum_{k=1}^{m} w_{1,k}^2 O_k^1 + b) \tag{9}$$

The neural network initializes the weights W^1 and W^2 with some random values. Hence the weights needs to be adjusted. Backpropagtion is a technique [5] in artificial neural network to adjust the weight of each nodes (neurons) of network by calculating the gradient of error function.

The error at node 1 of output layer (Class 1) is calculated using squared error [5] function E_{o1} represented by:

$$E_{o1} = 1/2(target - output)^2 \tag{10}$$

where, $target$ is the actual output and $output$ is the value which the node actually produce.

Similarly, error function E_{o2} and E_{o3} of node 2 (Class 2) and node 3 (Class 3) of output layer respectively should be calculated using Eq. 10.

Total error E_{total} of the network is represented by:

$$E_{total} = E_{o1} + E_{o2} + E_{o3} \tag{11}$$

Now the network will propagate backward to reduce the error by changing the values of weights. For that it needs to calculate the rate of change of error E_{total} with respect to change in weight represented by $\delta E_{total}/\delta w_{1,1}^2$:

$$\delta E_{total}/\delta w_{1,1}^2 = (\delta E_{total}/\delta O_1^2) \quad * \quad (\delta O_1^2/\delta n_1^2) \quad * \quad (\delta n_2^2/\delta w_{1,1}^2) \tag{12}$$

So, to adjust the weight $w_{1,1}^2$,$\delta E_{total}/\delta w_{1,1}^2$ is calculated and the weight$w_{1,1}^2$ is updated using following equations:

$$(w_{1,1}^2)^+ = w_{1,1}^2 - \epsilon * \delta E_{total}/\delta w_{1,1}^2 \tag{13}$$

Similarly, all other weights from hidden layer to output layer needs to be updated.

Also, the weights from input layer to hidden layer needs to be adjusted. To adjust the weight $w_{1,1}^1$, $\delta E_{total}/\delta w_{1,1}^1$ is calculated and the weight $w_{1,1}^1$ is updated by using following equ.:

$$\delta E_{total}/\delta w_{1,1}^1 = (\delta E_{total}/\delta O_1^1) \quad * \quad (\delta O_1^1/\delta n_1^1) \quad * \quad (\delta n_1^1/\delta w_{1,1}^1) \tag{14}$$

$$(w_{1,1}^1)^+ = w_{1,1}^1 - \epsilon * \delta E_{total}/\delta w_{1,1}^1 \tag{15}$$

Likewise, all other weights from input layer to hidden layer needs to be updated and this whole backpropagation process repeated five thousand times to minimize the error of the network. Once the neural network is designed, it needs to be trained with predefined or known traffic data set. The paper considers traffic density and average speed as input to the neural network and based on input neural network classifies the traffic congestion level. Why average speed along with density used as a input explained in result section. This paper considers three levels of congestion: high congestion, medium congestion and free flow.

4 SIMULATION AND RESULTS

To design and simulate the proposed artificial neural network based intelligent traffic congestion classification system, Python has been used.

The initial weights of connection between input layer nodes and hidden layer nodes are chosen randomly and the matrix is:

$$W^1 = \begin{bmatrix} -11.09 & -6.95 & -0.55 \\ 0.63 & 1.45 & -11.19 \\ -2.76 & -1.81 & 2.98 \\ 4.23 & 0.15 & 35.44 \\ 14.68 & -19.61 & -29.77 \end{bmatrix}$$

The initial weights of connection between hidden layer nodes and output layer nodes are also chosen randomly and the matrix is:

$$W^2 = \begin{bmatrix} -3.71 & -10.87 & 8.63 & -7.17 & 6.33 \\ 12.56 & -1.57 & 1.45 & -17.56 & 11.86 \\ -12.05 & -3.80 & 0.44 & -9.63 & -7.94 \end{bmatrix}$$

After training the proposed neural network five thousand times by backpropagation technique, the weights (W^1, W^2)are adjusted. The final weights of connection between input layer nodes and

hidden layer nodes are represented by the matrix:

$$W^1 = \begin{bmatrix} 0.26 & -1.56 & -0.09 \\ 0.05 & -0.06 & 1.84 \\ -1.99 & 0.80 & 0.84 \\ 0.80 & 1.62 & 0.32 \\ 0.95 & -0.24 & -0.08 \end{bmatrix}$$

The final weights of connection between hidden layer nodes and output layer nodes are represented by the matrix:

$$W^2 = \begin{bmatrix} -0.73 & -0.40 & -0.04 & -1.47 & 1.47 \\ 0.08 & -1.08 & 1.11 & 1.27 & 0.67 \\ -1.66 & 0.67 & 0.46 & 1.31 & 0.44 \end{bmatrix}$$

The road traffic information is fed into the ANN. For the simulation, two traffic parameters: traffic density (I_1) and average traffic speed (I_2) have been considered. These two parameters along with road segment (RS) identification number is considered as input to the neural network. The sample traffic parameters shown in Table 1 are captured through survey conducted on different area within the kolkata city and human perception is being used to categorize the congestion level. Three road segments RS1, RS2 and RS3 are considered and 40 set of inputs for each road segments have been used.

Before the traffic parameters are fed into neural network, scaling is performed. Different road segment identification numbers (RS) are transformed into numeric value before feeding into the neural network. The neural network needs to be trained with some known traffic data set so that it can accurately classify the traffic congestion level of different road segments with respect to density and average speed of vehicles. As it is considered three congestion level high, medium and free flow, the proposed network have three output each corresponding to different congestion level. The output is encoded in numeric value ranging between zero and one with the help of sigmoid activation function mentioned in Eq. 5. The scaled parameters along with corresponding output is shown in Table 2.

The proposed neural network is trained with traffic data set mentioned in Table 2 for different types of road segments with varying lane width.

Once the network is trained, it can identify traffic congestion level of a road segment based on density and average speed. Some of the predicted traffic congestion level based on density and average speed is shown in Table 3. It shows that the proposed neural network model accurately classify the traffic congestion state. To justify the accuracy of the proposed system, field survey has been done on different road segments in Kolkata city. For example, if road traffic density is 340 vehicles/km and average speed 20 km/hr. of road segment 1 (RS1) (the scaled value is shown in Table 3, traffic density = 3.4 vehicles/km and average speed = 0.2 km/hr.), then proposed neural model categorizes it as a class 1 (high congestion) which perfectly matches with human observation (i.e. from field survey also it has been observed as high congestion in road segment (RS1) for the above mentioned scenario). Similarly, if road traffic density is 100 vehicles/km and average speed 80 km/hr. of road segment 1 (RS1), then proposed neural model categorizes it as a class 3 (free flow). As different road segments are of different types with varying lane width, therefore some road segments can become congested even when lesser number of vehicles is present in the road. In these cases also, the proposed ANN based model can classify the congestion states accurately. For example, in road segment 2 (RS2) which is narrower than RS1, if road traffic density is 155 vehicles/km and average speed 17 km/hr., then it categorize it as a class 1 (high congestion) despite of presence of lesser number of vehicles than RS1. The proposed ANN model identifies road segment 3 (RS3) as a high congestion states when it has 220 vehicles/km and average speed is 18 km/hr., which is also accurately matches with on field human observation. The congestion status in RS3 is identified as a medium congestion when it has 90 vehicles/km and average speed is 52 km/hr. as the scaled value shown in Table 3.

The proposed system is compared with W. Pattara-atikom et al. [12] with respect to human observation to evaluate the accuracy of the system. The comparison has been done on the dataset shown in Table 1. In [12], the authors considered only vehicle velocity to estimate traffic congestion. They used two threshold values β

Table 1: Sample Traffic Parameters

Road Segment Identification	Traffic Parameter	
	Density (No. of cars/km)	Avg. Speed (km/hr.)
RS1	380	12
RS1	350	18
RS1	320	23
RS1	300	28
RS1	270	39
RS1	255	42
RS1	248	48
RS1	210	55
RS1	190	60
RS1	150	72
RS1	130	78
RS1	90	82
RS1	60	90
.	.	.
.	.	.
RS2	180	12
RS2	150	18
RS2	120	25
RS2	100	35
RS2	80	42
RS2	70	47
RS2	62	53
RS2	48	70
RS2	10	90
.	.	.
.	.	.
RS3	250	16
RS3	230	16
RS3	180	27
RS3	120	43
RS3	100	50
RS3	65	75
RS3	50	77

Table 2: Sample Training Data Set

Training Data Set					
Road Segment Identity	Scaled Traffic Parameter		Output		
	Density (No. of Car/km)	Avg. Speed (km/hr.)	Class1	Class2	Class3
1	3.8	0.12	1	0	0
1	3.2	0.23	1	0	0
1	2.7	0.39	1	0	0
1	2.55	0.42	0	1	0
1	2.1	0.55	0	1	0
1	1.9	0.60	0	0	1
1	1.3	0.78	0	0	1
1	0.6	0.90	0	0	1
.	.	.	.	.	.
.	.	.	.	.	.
2	1.8	0.12	1	0	0
2	1.2	0.25	1	0	0
2	1.0	0.35	1	0	0
2	0.8	0.42	0	1	0
2	0.62	0.53	0	1	0
2	0.30	0.70	0	0	1
2	0.10	0.90	0	0	1
.	.	.	.	.	.
.	.	.	.	.	.
3	2.50	0.16	1	0	0
3	2.30	0.16	1	0	0
3	1.80	0.27	1	0	0
3	1.20	0.43	0	1	0
3	1.00	0.50	0	1	0
3	0.65	0.75	0	0	1
3	0.50	0.77	0	0	1

and γ to classify the congestion levels into three categories: red, yellow and green which correspond to high congestion, medium congestion and free flow levels respectively. If average velocity is less than β km/hr, then the system classify it as a red level i.e. high congestion while it classifies as a yellow level if the average velocity is between β km/hr and γ km/hr, and when the average velocity is greater than γ, the system classify the road segment as green level i.e. free flow. However the system does not always classify the traffic congestion levels accurately. Only vehicle velocity cannot be considered as a measure to classify congestion levels. This is because velocity may be less than threshold value β not due to congestion but due to bad road condition or weather condition (like heavy rainfall, fog etc). The intelligent traffic congestion classification system presented in this paper considers traffic density and average speed to overcome the above mentioned problem presented in [12]. Thus, the proposed work classifies the congestion level accurately. Figure 3 shows the comparison between the proposed system and W. Pattara-atikom et al. system in terms of accurate classification of congestion level. It uses the users' observations for classifying congestion level and it is shown in the figure as the reference line. High congestion is represented by 1 while medium congestion and free flow is represented by 2 and 3 respectively in the comparison graph.

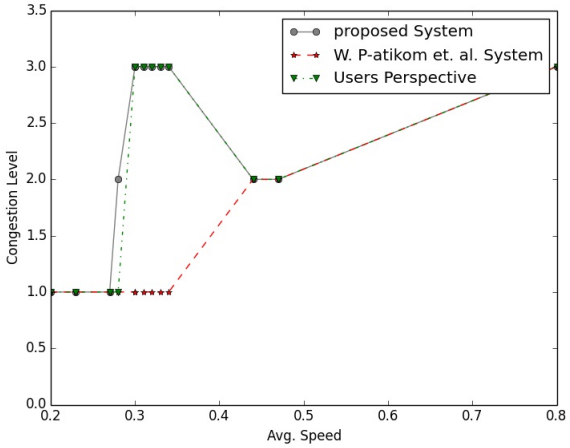


Figure 3: Performance analysis with respect to average speed and congestion level.

Table 3: Sample Traffic Congestion State Prediction

Training Data Set					
Road Segment Identity	Scaled Traffic Parameter		Output		
	Density (No. of Car/km)	Avg. Speed (km/hr.)	Class1	Class2	Class3
1	3.40	0.20	9.99e-01	3.12e-05	6.08e-04
1	3.0	0.23	9.99e-01	5.48e-05	5.85e-04
1	2.9	0.27	9.99e-01	6.60e-05	4.99e-04
1	2.55	0.44	5.15e-03	9.93e-01	2.09e-03
1	2.5	0.47	2.04e-04	9.98e-01	8.62e-04
1	1.0	0.80	3.62e-10	1.11e-03	9.98e-01
1	0.5	0.34	2.04e-06	1.14e-01	8.90e-01
1	0.48	0.33	2.78e-06	1.91e-01	8.08e-1
1	0.4	0.31	4.03e-06	2.95e-01	6.86e-01
1	0.39	0.32	2.32e-06	1.23e-01	8.77e-01
1	0.38	0.3	5.77e-06	4.34e-01	5.24e-01
1	0.35	0.3	4.56e-06	3.19e-01	6.51e-01
1	0.35	0.28	6.08e-5	8.78e-01	3.23e-01
2	1.35	0.17	9.99e-01	3.21e-05	5.88e-04
2	0.68	0.48	4.03e-05	9.99e-01	1.22e-03
2	0.25	0.78	3.99e-10	1.11e-03	9.98e-01
3	2.20	0.18	9.99e-01	6.55e-04	6.77e-08
3	2.00	0.19	9.98e-01	3.09e-03	8.47e-08
3	0.90	0.52	1.74e-01	8.08e-01	3.17e-08
3	0.60	0.72	5.56e-06	3.65e-02	9.52e-01

The graph shown in Figure 3 is drawn according to traffic data presented in Table 3. The paper sets the value of β with 35 km/hr. and value of γ with 55km/hr. Based on these threshold values,

whenever the average speed is below 35km/hr. the W. Pattara-atikom et al. system always classified the road segments as a highly congested road. But it does not always reflect the actual scenario for example during adverse weather conditions (such as during heavy rainfall, fog) or due to poor road condition the aver-age vehicular speed might be decreased irrespective of the number of vehicles present in the road. For the data sets of table 3: (1, 0.5, .34), (1, 0.48, .33), (1, 0.4, .31),(1, 0.39, .32), (1, 0.38, .3), (1, 0.35, .3),(1, 0.35, .28), W. Pattara-atikom et. al. system wrongly classifies the congestion level as a high congestion level. But the proposed system classifies the congestion level accurately for the above mention data sets as indicate in Table 3. The proposed ANN based system classifies the congestion level as free flow even when the average speed is less than threshold value (35km/hr.) taking into account the value of traffic density as well which almost matches with the result obtained from field survey.

5 CONCLUSION

This paper proposes an intelligent traffic congestion state classification system based on artificial neural network model (ANN). Here we consider two road traffic parameters: traffic density and speed as measures to classify traffic congestion states and three congestion levels: high congestion, medium congestion and free flow. ANN model is designed to determine the traffic congestion status. After training the neural network, it determines the congestion status of road segments. From the result it is seen that the proposed system accurately classifies the congestion status for different types of road segments with varying lane width provided the model is perfectly trained with adequate traffic data sets. Based on traffic congestion status, ITS will automatically update the traffic regulations like, changing the queue length in traffic signal, suggesting alternate routes. It also helps the government to device policies regarding construction of flyover/alternate route for better traffic management.

REFERENCES

[1] L. Anthopoulos, M. Janssen, and V. Weerakkody. 2016. A Unified Smart City Model USCM for Smart City Conceptualization and Benchmarking. *Int. J. Electron. Gov. Res.* 12, 2 (April 2016), 77–93. https://doi.org/10.4018/IJEGR.2016040105

[2] T. Cheng, P. Wen, and Y. Li. 2016. Research Status of Artificial Neural Network and Its Application Assumption in Aviation. In *2016 12th International Conference on Computational Intelligence and Security (CIS)*. 407–410. https://doi.org/10.1109/CIS.2016.0099

[3] Benjamin Coifman and SeoungBum Kim. 2009. Measuring Freeway Traffic Conditions with Transit Vehicles. *Transportation Research Record* 2121, 1 (2009), 90–101. https://doi.org/10.3141/2121-10 arXiv:https://doi.org/10.3141/2121-10

[4] Wiam Elleuch, Ali Wali, and Adel M. Alimi. 2017. Intelligent Traffic Congestion Prediction System Based on ANN and Decision Tree Using Big GPS Traces. In *Intelligent Systems Design and Applications*, Ana Maria Madureira, Ajith Abraham, Dorabela Gamboa, and Paulo Novais (Eds.). Springer International Publishing, Cham, 478–487.

[5] Martin T. Hagan, Howard B. Demuth, and Mark Beale. 1996. *Neural Network Design*. PWS Publishing Co., Boston, MA, USA.

[6] Simon Haykin. 2009. *Neural Networks And Learning Machines* (third edition ed.). Pearson Education Inc, Upper Saddle River. Includes Index.

[7] Feifei He, Xuedong Yan, Yang Liu, and Lu Ma. 2016. A Traffic Congestion Assessment Method for Urban Road Networks Based on Speed Performance Index. *Procedia Engineering* 137 (2016), 425 – 433. https://doi.org/10.1016/j.proeng.2016.01.277 Green Intelligent Transportation System and Safety.

[8] B. Krause, C. von Altrock, and M. Pozybill. 1996. Intelligent highway by fuzzy logic: congestion detection and traffic control on multi-lane roads with variable road signs. In *Proceedings of IEEE 5th International Fuzzy Systems*, Vol. 3. 1832–1837 vol.3. https://doi.org/10.1109/FUZZY.1996.552649

[9] J. Lu and L. Cao. 2003. Congestion evaluation from traffic flow information based on fuzzy logic. In *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, Vol. 1. 50–53 vol.1. https://doi.org/10.1109/ITSC.2003.1251919

[10] Andres Monzon. 2015. Smart Cities Concept and Challenges: Bases for the Assessment of Smart City Projects. In *Smart Cities, Green Technologies, and Intelligent Transport Systems*, Markus Helfert, Karl-Heinz Krempels, Cornel Klein, Brian Donellan, and Oleg Guiskhin (Eds.). Springer International Publishing, Cham, 17–31.

[11] Hasan Omar Al-Sakran. 2015. Intelligent Traffic Information System Based on Integration of Internet of Things and Agent Technology. *International Journal of Advanced Computer Science and Applications* 6, 2 (2015). https://doi.org/10.14569/IJACSA.2015.060206

[12] W. Pattara-atikom, P. Pongpaibool, and S. Thajchayapong. 2006. Estimating Road Traffic Congestion using Vehicle Velocity. In *2006 6th International Conference on ITS Telecommunications*. 1001–1004. https://doi.org/10.1109/ITST.2006.288722

[13] F. Porikli and X. Li. 2004. Traffic congestion estimation using HMM models without vehicle tracking. In *IEEE Intelligent Vehicles Symposium, 2004.* 188–193. https://doi.org/10.1109/IVS.2004.1336379

[14] T. Rawal and V. Devadas. 2015. Intelligent Transportation System in India- A Review. *Journal of Development Management and Communication* 2, 3 (April 2015).

[15] Zeenat Rehena and Marijn Janssen. 2018. Towards a Framework for Context-Aware Intelligent Traffic Management System in Smart Cities. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 893–898. https://doi.org/10.1145/3184558.3191514

[16] Sten Ruppe, Marek Junghans, Mathias Haberjahn, and Christian Troppenz. 2012. Augmenting the Floating Car Data Approach by Dynamic Indirect Traffic Detection. *Procedia - Social and Behavioral Sciences* 48 (2012), 1525 – 1534. https://doi.org/10.1016/j.sbspro.2012.06.1128 Transport Research Arena 2012.

[17] Bhupendra Singh and Ankit Gupta. 2015. Recent trends in intelligent transportation systems: a review. *Journal of Transport Literature* 9 (04 2015), 30 – 34. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2238-10312015000200030&nrm=iso

[18] Abhinav Sunderrajan, Vaisagh Viswanathan, Wentong Cai, and Alois Knoll. 2016. Traffic State Estimation Using Floating Car Data. *Procedia Computer Science* 80 (2016), 2008 – 2018. https://doi.org/10.1016/j.procs.2016.05.521 International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

[19] Thammasak Thianniwet, Satidchoke Phosaard, and Wasan Pattara-Atikom. 2009. Classification of Road Traffic Congestion Levels from GPS Data using a Decision Tree Algorithm and Sliding Windows.

Leveraging AI and Sensor Fabrics to Evolve Smart City Solution Designs

William H. Money
The Citadel
Baker School of Business, Charleston,
South Carolina, United States,
willmny@gmail.com

Stephen Cohen
Azure Global
Microsoft Corp., Washington DC,
United States, stcohen@microsoft.com

ABSTRACT

Cities have entered the age of the sensor and located sensors everywhere over and under cities. The sensors monitor a host of factors that assess City operations and life such as air quality, noise, city services and traffic. Further, the sensors have "gone mobile" with announcements of situation aware mobile sensor platforms designed for city-level security and public safety. These wearable sensor platforms combine video, audio, and location data with Internet of Things (IoT) capabilities. However, the many sensors and functional platforms have not yet made the cities employing these many diverse sensors truly Smart. We are analyzing why the success toward the Smart city is limited, or late in coming. The explanations for the constrained effectiveness are assigned to many factors, but one of significance can be teased from a long-accepted explanation that associates data, information, and knowledge. Smart Cities need to effectively use the sensor data and the information assembled from these interpreted and organized data to create knowledge that serves the city and its people by answering and resolving key problems and questions. But the systems and analytic models needed to associate these data from many sensors have yet to be designed, constructed, and proven in the complex cities of today. Thus, the data (and information from the diverse sensors) lacks crucial integration and coordination for decisions and sense-making. While these sensor-based systems were, and in many cases are meeting some intended functionally discrete goals, they appear to be better described as data collection tools feeding centralized analytical engines. They are point solutions with specialized or targeted sensors feeding specialized solutions. This is a significant limiting factor in a city's drive to improve the quality of life and the efficiency of the services a city provides to its stakeholders. In this paper we present current trends in Smart City development, emerging issues with data and complexity growth, and proposes a mean to leverage the advancing technologies to address the integration problem.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317049

CCS CONCEPTS

• Computing methodologies~Reasoning about belief and knowledge • Computing methodologies~Reasoning about belief and knowledge;500 • Computing methodologies~Learning from implicit feedback • Computing methodologies~Reasoning about belief and knowledge;300 • Computing methodologies~Cognitive science • Computing methodologies~Learning from implicit feedback;300

KEYWORDS

Smart City, Artificial Intelligence, Sensors, Sensor Fabric

ACM Reference format:

William H. Money and Stephen Cohen. 2019. Leveraging AI and Sensor Fabrics to Evolve Smart City Solution Designs. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317049

1 Introduction

The concept of the Smart City is one of the major trends that appears to generate strong and continuing research and discussions as we have moved well into the 21st century. Humankind seems to be moving into rather that away from urban environments. This stimulates concentrated growth and requires the development of new technologies, processes and techniques for managing the complexity of metropolitan centers. One approach to address this is to employ sensors and collect data to support the proactive management of cities functions. The number and variety of Smart alternatives, creative ideas for sensors, deployment of mobile devices, and current sources of tested and operative sensor data are truly impressive. They include:

- carpets of sensors across a City for autonomous cars (driverless cars),
- apps car drivers crossing over a pothole,
- digital masterplans for City designs,
- structure sensors for sustainable buildings,
- landing posts (utility poles) connected to underground, optical wiring that could provide neighborhoods with LED street lighting,
- process parking transactions, and

- service and data lists serving as electronic neighborhood bulletin boards. [1]

Added to these devices are end-to-end platforms supporting the aggregation and correlation of near real and real time data streams. Mobile sensor platforms designed for city-level security and public safety operations, such as wearable sensors, combine video, audio, location data, and Internet of Things (IoT) capabilities, which can be centrally managed via cloud-based software to offer real-time "situational awareness". One example is a public safety solution by Flir Systems which provides a system of sensors that can be worn on an individual's body or mounted inside vehicles. It has features that deliver data from visible-video, audio, global navigation satellite system (GNSS), gyroscope, accelerometer and magnetometer sensors. This provides decision makers more information so they can respond to complex situations and evolving threats as they occur [2].

2 Sensor Objectives and Purposes

It is critical that Smart cities apply the swarms of available sensors wisely and strategically to benefit citizens with more rapid and higher quality outcomes. If Smart cities fail to provide clear social and economic benefits, the point of the sensor-managed city would become merely to serve as a revenue generating source for the various commercial organizations providing and managing the technology.

To meet the increasing need for innovative solutions to the problems of the urban environment, Smart City data and the resulting analytical outcomes from the sensors must consistently influence ongoing efforts to optimize overall city performance. This requires planning (for services like preventive maintenance, security, traffic flows…etc.), actual delivery of the services, and post-delivery monitoring. These services must be performed with a clear citizen-focused goals that minimize the costs of services and maximize the quantity and quality of the delivery of the services. This structured and coordinated purposeful decision making has been described and envisioned by cities and Smart city researchers for many years. Yet the underlying highly coordinated and qualitatively effective system has not evolved in a significant way. Since 2000 the vision has been of a city where an urban center is "…safe, secure environmentally green, and efficient because all structures--whether for power, water, transportation, etc. are designed, constructed, and maintained making use of advanced, integrated materials, sensors, electronics, and networks…" This city image predicts that the Smart City will be managed by computerized systems comprised of databases, tracking, and decision-making algorithms [3].

While the initial vision called for systems to be functionally focused or tailored for an application, it also called for a system with a design that would;

- make use of these building-block like components,
- provide an interface to the computerized "monitoring" capability for each given function, and
- have a full structure or service supplied.

Finally, the vision posited the system would somehow be integrated across all aspects of an urban center's essential infrastructure. The envisioned capability of these all-encompassing macro systems was that they would monitor and integrate all "…critical infrastructures, including roads, bridges, tunnels, rail/subways, airports, seaports, communications, water, power, and major buildings." Proponents and advocates of Smart cities asserted and often predicted that systems and structures would have the capacity monitor their own conditions and carry out self-repair. However, it was never clear how this transformation would occur. Even the physical environment including air, water, and surrounding green spaces will be monitored in non-obtrusive ways to maintain optimal quality levels [3].

The limiting factors in this early period of Smart City sensor application development recognized the limitation of sensors with embedded microprocessors and wireless communication links. A 2002 National Research Council report discussed the benefit of networked systems of embedded computers and sensors throughout society that could provide support for a greatly expanded information revolution. The report noted that a framework did not then (~2004) exist that could allow the distributed computing paradigm offered by smart sensors to be beneficially applied to structural health monitoring and control systems. The report cited the Tsing Ma Bridge monitoring system established in Hong Kong in 1997 with 280 sensors, including anemometers, temperature sensors, strain gauges, accelerometers, global positioning systems (GPS), displacement transducers, and level sensors as well as the Stonecutters Bridge monitoring system (Hong Kong) with over 1500 sensors to support the contention. The problem recognized was the algorithms in use assumed that all data would be centrally collected and processed. It was noted that the approach did not scale to systems with densely instrumented arrays of sensors required, and that practical and durable power supplies were still a challenge [4].

2.1 Today, and the Future of Sensor Data

It is important to comprehend the scope and scale of the sensor driven Smart City to fully appreciate the problem facing managers, system architects, data managers, and a host of others involved in making cities Smart. First, we need to catalog what sensors inhabit the Smart City environment, how much data is produced, and what it takes to process the data even before we try to make sense of the data and perform coordinated correlation and analysis activities. There are many ways of exploring and categorizing the Smart City sensors of today [5]. They encompass assessments of the breadth of the sensors placed around a City, the functional diversity of the sensors, the data generated and available from the sensors and the time sensitivity and locational volatility of mobile sensors.

As a starting point, consider the City of Barcelona which has been described as a city that can be categorized as being very "Smart." [6] Barcelona encompasses about 100 Km^2 and has about 1.62 million inhabitants. Functionally, city owned equipment includes 150,000 lampposts, 40,000 garbage containers, and 80,000 public parking spots. The cities data are collected by a sensor networks for energy monitoring, Noise

monitoring, urban environmental monitoring (includes air, temperature and humidity measurements), garbage collection monitoring, and parking spots monitoring are all in operation. [6] The sensor data generated may be viewed by the amount, type, and frequency of the generated data. These data are then combined to obtain data estimates of daily sensor traffic [7].

This city has almost 1,800 sensors installed monitoring these energy, noise, urban lab, garbage collection, and parking at specific locations. Energy monitoring is conducted for municipal buildings and solar thermal installations. It provides energy consumption (electriCity meter, electriCity ambient conditions, gas meter, internal ambient conditions, and temperature) while the solar thermal installation ascertains solar thermal energy produced and consumed. Noise monitoring management detects noise and acoustic pollution (about 50 distributed sensors) in seven different city areas. Laboratory environmental monitoring assesses air, temperature, humidity, and some transportation issues (including people and bicycle flow) with about 50 active sensors spread through the various zones of Barcelona. Garbage Collection management obtains data about garbage and containers with hanging sensors. The refuse data are organized by glass, organic, paper, plastic and refuse through about 660 sensors. Parking Spots monitoring locates free parking spaces through approximately 500 sensors in one localized area of Barcelona. The monitoring service transfers and updates several data packets periodically during the day.

The resulting data generated per day are:

- Energy monitoring management (>3MB),
- Noise monitoring management (578KB),
- Urban Lab monitoring management (153KB),
- Garbage Collection management (480KB), and
- Parking Spots management (615KB).

About 5MB that is transferred through the Barcelona Smart City network. Barcelonas' data are particularly useful because they can be analyzed and extrapolated to estimate the generated data from similar services for an entire city. According to Barcelona's City statistics, there were 70,000 buildings, 40,000 containers and 80,000 parking slots in 2014. Sinaeepourfard et. al [7] estimated that there are 40,000 street corners. The number of sensors, and corresponding expected generated data (for a city-wide equivalent network) is about 8 GB per day from all sensors in the city. (This includes only data obtained from these identified sensors). The number of sensors needed would come to 321 million sensors for coverage of the whole city. Excluded are data from mobile devices, surveillance cameras, information from web services, etc [7].

3 Stationary and Dynamic Sensors and sensor platforms

Fast forward to today and we find ourselves entering a new era of robots with great capabilities — flying drones functioning in both natural and man-made environments. Initially drones were associated with defense applications but they are having an increasingly significant impact on civilian tasks such as transportation, communication, agriculture, disaster mitigation and environment preservation. The determinants of the drone effectiveness are due in part to the energy required to staying airborne, the perceptual intelligence required to negotiate complex environments, and the tools available for collection and reporting of data that may be useful. It has been predicted that there will be great use of autonomous drones for civilian applications [8].

Drones have significant functionality for diverse products including commercial inspections, photogrammetry and movie making. But drone makers have found it necessary to continuously improve and increase the number and capability of sensors, so they possess features that are necessarily protective of the sensing (drone) vehicle itself. Today consumer and professional drones are being constructed with obstacle detection and collision avoidance sensors. This technology began with drones equipped with sensors detecting objects in front of the drone and has progressed to drones from DJI, Walkera, Yuneec and others have front, back, below and side obstacle avoidance sensors including Stereo Vision, Monocular Vision, Ultrasonic, Infrared, Time-of-Flight and Lidar sensors (with only one drone with all 6 directions of obstacle detection today). To understand detected objects and then take action to avoid the obstacle - whether to stop, go around or above the object, is difficult manually. The operator must recognize and then take immediate and appropriate actions. Operating an autonomous drone require is more difficulty and requires an array of obstacle detection sensors including SLAM technology to interpret the images being scanned (by the sensors), GPS and GLONASS satellite navigation systems to know exactly where it is and maintain stable flight as well as software algorithms and software programming which includes mathematical modelling, algorithms, machine learning [9].

The resulting and notably impressive sensor technology of the drone generates potentially vast amounts of data. Functioning together, they include:

- Gyros and accelerometers (tilt controls) to determine position and orientation of the drone in flight.
- thermal sensing of changes in the movement of gas molecules passing over a small integrated circuit on-board camera
- Inertial measurement units
- GPS utilize multi-axis magnetometers
- Current sensors monitor and optimize power drain, detect fault, and safely charge internal batteries, conditions with motors or other areas of the system
- Gas engine mass-flow sensors employ an calorimetric principal utilizing a heated element to monitor air flow into small gas engines to maintain the proper fuel-to-air ratio at a specified engine speed.

All these sensors feed data into a central processor to maintain adherence to flight rules, air traffic control direction, and flight paths [10].

4 Inbuilt intelligence, self-diagnostics and repair.

The manufacturing response from sensor makers appears clear. The focus has been on the sensor itself. For example, the sensor community has begun to focus on built in intelligence. It is essential to reduce the huge volumes of generated data into actionable information for users, and commands given to actuators back out in the field. Sensors which simply convert physical variables into electrical signals must evolve into something more sophisticated to perform a technically and economically viable role. Manufacturers have responded with improvements to fabrication, adding integration and built-in intelligence. This facilitates IoT connectivity, improves predictive maintenance, enables self-diagnostics and repair and improves manufacturing. Manufactures have also improved the base functionality by sending calibration data to the MPU so that the sensor is automatically set up for any production changes and can identify production parameters that start to drift beyond acceptable norms. It can then generate warnings so preventative action can be taken.

This 'report by exception' reduces both the load on the central computing resource, and the smart sensor's power requirements. Dual elements enable sensor self-diagnostics since drift in one of the sensor element outputs can be detected or the process can continue with the second measuring element. A Texas Instruments product example uses an ultra-low-power sensor to build a smart fault indicator for electric power distribution networks. The fault indicators reduce operating costs and service interruptions by providing information about a failed section of the network. Fault indicators, which are installed on the junctions of the overhead power-line network, send measurement data for power transmission lines wirelessly to the concentrator/terminal units mounted on the poles which then passes the data to the cellular network to relay real-time information to the main station.

The resulting Intelligent sensors can be remotely programmed with suitable parameters every time a product change is required. Sensors can self-calibrate by using a patented combination of an ASIC and an array of MR (magneto resistive) sensors. This accurately and reliably determines the position of a magnet attached to moving objects such as elevators, valves or machinery. The output and the MR sensor sequence determine the nearest pair of sensors to the center of the magnet location. The output from this pair is then used to determine the position of the magnet between them. This non-contacting self-diagnostics feature technology can provide enhanced product life, reduce downtime levels, and increased durability. Similar manufacturing goals to integrate all the elements into a single discrete device and adding intelligence and communications capabilities to the basic device provides further evidence of manufactures continuing response to IoT's needs [9].

5 Challenges using Smart City sensor data

Beneficial and game changing uses of the Smart city data requires more than just additional sensors and enormous data streams. As the evolution of drones shows, sensors must work in a coordinated and strategic manner to address deep problems in Smart cities. This effectiveness can be achieved by constructing a fabric that unifies a city. Such a fabric will require both organizational changes addressing city decision making and technical developments. Some suggestions for organizational and technical (but not combined) Smart City solutions must include technical architectural models such as the IBM Smart Cities technical architecture which helps to making sense of the data; not as previously assumed, people do not only want efficiencies but seek differentiation and some way of attracting Smart talent to create economic development [1]. Providing a consistent fabric of city management presents the dual challenge of technical complexity and legislative and ethical controls.

City managers must also deal with the impedance mismatch between the rate of change in technological capability and the legislative and legal systems ability to adapt to the availability of new technologies. Consider the long history of collecting sensor information from remote flying vehicles that can be traced back to the use of satellites and the limitation placed on these data by U.S laws. Images and information available from remote-sensing satellites was potentially valuable to both commercial and news media. Ready to make major use of new near real-time aerial and satellite information (which in the early 1980s offered resolution of less than 10 meters), potential consumers had to deal with licensing restrictions placed on any US company wanting to put a camera into orbit by the US Land Remote Sensing Commercialization Act of 1984 [11]. Technical capability blocked by legislative control.

The technical challenge for Smart Cities is to find ways of combining streaming and collected data with Smart City performance goals, concepts, and architectures. The progress assessment for this high level of coordination among the various architectures and their related data do not provide yet any comprehensive model which allows for consistently integrating all expected Smart Cities functionalities. Currently, things such as accepted standardization or data collection globalization are limited or simply do not exist. In addition, the increasing speed of developments in technology and innovation to Smart city data sources challenges city managers to consume and analyze data while citizens and stakeholders demand immediate service delivery and benefits [3].

6 Integration and Fabric Development Baby Steps

Most smart cities are evolving the underlying infrastructure as they deploy more services and build collaborative integration. Smart city integration of services requires highly reliable, highly performant networks, both fiber and wireless, as well as elastic compute and storage. Such complex tools are difficult to develop and slow to implement. Solutions that can more rapidly advance the integration required are necessary, if they exist. This paper suggests that the need to implement and enforce governance across all deployed solutions and services can be more rapidly

addressed if the Smart cities create and provide or more likely procure use of a Cloud Fabric.

Service Fabric itself is an application platform layer. As the hosting platform the fabric provides an abstraction layer over the underlying infrastructure, helps manage the lifecycle of deployed services, and removes most external dependencies while unifying the deployment, management, and upgrade models [12]. The fabric host is a virtual integration space as long as it provides continuous monitoring; records the underlying metrics of operational use; delivers notifications required by the hosted solution and hosting system's status and health; and may be used to provided mandatory, certified components supporting a unified governance model.

It is assumed that there will be no "earthquake–like" fabric development and deployment model. The fabric (or multiple smaller focused fabrics) will grow or evolve through prototyping, agile integration of Smart development projects, and coordinated devops solutions. Once the integration challenges have been effectively addressed, Smart Cities will be able to confront the volume and speed of the real time data sensors continuously produce.

7 The future of AI is Assistive, not Artificial, Intelligence

Decision makers across all role of city management rely on available information to make their decisions. The introduction and growth of sensors and the proliferation of available data can strain available information handling processes. The sheer volume, as discussed above, may obfuscate or overwhelm the demands of the integration effort, and its context. We believe that Smart city systems will likely evolve in several stages beginning with functionally successful stand-alone solutions. We the expect to see a development of more tightly integrated automated systems. We then expect to encounter cities that incorporate collaborative sensor networks, and finally we see the city systems evolving into autonomous and assistive tools that well manage the Smart cities of our future. Obviously, this evolutionary framework can be altered by the strategy and relative benefits (and needs) of the functionality being introduced in a specific urban center.

The following discussion helps us better understand and visualize what must be accomplished by these systems. For example, consider at the all too common scenario; when an unaccompanied driver is attempting to rapidly enter a busy roadway.

In a city using highly **integrated** technologies and operator may be provided video alerts due to movement within a predefined zone. They might receive the density and speed data for the roadway traffic and be able to make a rapid adjustment to the road signals to slow if not stop traffic. Today these independent systems are integrated by the driver/user. Providing the situational awareness to the operator is a significant improvement to previously fielded 'dumb' systems.

If we could create a more **collaborative** environment by hosting the individual system components on a shared fabric the operator would not need to seek secondary systems. The video detection would proactively display nearby, managed, systems. It could share location and state of those systems and likely reduce the speed so an operator could react to the potential danger.

An **autonomous** system would be permitted to take some additional action within the bounds of governance, policy, and law. Detecting the motion path and potential threat to the person the autonomous system might proactive change nearby stop lights to yellow to slow ongoing traffic while alerting the operator in parallel with the change. The operator may choose to override the autonomous choice or may ask the autonomous solution to provide options before either the system or the operator takes action.

Ultimately, Smart cites can evolve their systems to become **assistive** partners in the management of the city. The great ability is that a cloud fabric-based solution can to act in parallel. It could consider the alternatives available and diverse needs of multiple and different users. What is required is that the cloud fabric-based solution consider an exceptionally large array of alternatives in a relatively small amount of time, and that it correlate a large historical model of past behaviors to provide decision makers proactive suggestions. The solution migh take note of consistent pedestrian motion paths and consider a history of pedestrian-vehicle accidents. Finally, it might suggest adding fencing to block foot traffic) as well as reactive alerting for the passing traffic.

We believe the next generation of Smart city assisted intelligence and decision support systems will leverage data science, machine learning, and context awareness. It will both reactively and/or proactively provide decision makers statistical analysis supported by machine learning. The systems will predict, engage, and prevent outcomes (as determined by the Smart city stakeholders and decision makers) while leveraging a full feedback loop to continuously evaluate and improve performance.

8 Conclusions

The balancing and allocation of benefits of the Smart city among the stakeholder parties is an important decision that can lead to increasing or decreasing total cost (and benefit) of a specific project. We do not propose that the systems make arbitrary decisions. Such decision can affect the overall relationship between the Smart city inhabitants. Common benefit assessment processes are based on experiential knowledge and are considered subjective or perhaps simply implicit and "good." Smart cities will require fuzzy adaptive decision making models for selection of benefit allocations which transforms the linguistic principles, likes and preferences, and experiential expert knowledge into a more usable and systematic quantitative-based analysis. Cities may use fuzzy logic qualitative approaches and analytic hierarchy process (AHP) adaptive capabilities to evaluate the allocation of benefits and determine where and when to distribute each benefit a specific use or user group to achieve the highest overall level of benefits.

This paper has discussed an explanation for the constrained effectiveness of the Smart city revolution. It argues that unassociated (stove-piped) data, information, and knowledge are limiting the effectiveness of the Smart city of today. Smart Cities need to use the sensor data and the information assembled from these interpreted and organized data to create knowledge that serves the city and its people by answering and resolving key problems and questions. But the systems and analytic models needed to associate these data from many sensors have yet to be designed, constructed, and proven in the complex cities of today. Thus, the data (and information from the diverse sensors) lacks crucial integration and coordination for decisions and sense-making. While these sensor-based systems were, and in many cases are meeting some intended functionally discrete goals, they appear to be better described as data collection tools feeding centralized analytical engines. They are point solutions with specialized or targeted sensors feeding specialized solutions. The paper concludes that a fabric to integrate the sensors and data for the Smart city can resolve this problem

REFERENCES

[1] Dennis, Roger. "Observations from presenting at the Nobel Cause Symposium," April 28, 2015. http://www.rogerdennis.com/ideaport/?cat=21 Last accessed January 20, 2019.

[2] "Mobile sensor platform announced for smart cities." SmartCitiesWorld, 14 Nov 2018. https://www.smartcitiesworld.net/news/news/mobile-sensor-platform-announced-for-smart-cities-3555 Last accessed, January 25, 2018.

[3] Hall, R. E., Bowerman, B., Braverman, J., Taylor, J., Todosow, H., & Von Wimmersperg, U. (2000). The vision of a Smart City(No. BNL-67902; 04042). Brookhaven National Lab., Upton, NY (US).

[4] The Smartest Cities In The World In 2018. Forbes. Jul 13, 2018. https://www.forbes.com/sites/iese/2018/07/13/the-Smartest-cities-in-the-world-in-2018/#7e2ae1f72efc Last accessed. January 21, 2019.

[5] Spencer Jr, B. F., Ruiz-Sandoval, M. E., & Kurata, N. (2004). Smart sensing technology: opportunities and challenges. Structural Control and Health Monitoring, 11(4), 349-368.

[6] C. Kyriazopoulou, "Smart City technologies and architectures: A literature review," in Smart Cities and Green ICT Systems (SMARTGREENS), 2015 International Conference on, 2015, pp. 1-12.

[7] Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., Marín-Tordera, E., Cirera, J., Grau, G., & Casaus, F. (2016, June). Estimating Smart City sensors data generation. In Ad Hoc Networking Workshop (Med-Hoc-Net), 2016 Mediterranean (pp. 1-8). IEEE.

[8] Banerjee, Ankit. Aug 16 2018. "Samsung SmartThings — everything you need to know." https://dgit.com/samsung-smartthings-57827/ (Las t accessed January 26, 2019)

[9] Floreano, D., & Wood, R. J. (2015). Science, technology and the future of small autonomous drones. Nature, 521(7553), 460.

[10] Corrigan, Fintan, December 28, 2018. "12 Top Collision Avoidance Drones And Obstacle Detection Explained." https://www.dronezon.com/learn-about-drones-quadcopters/top-drones-with-obstacle-detection-collision-avoidance-sensors-explained/ (Last accessed January 26, 2019)

[11] Farnell element14, "Smart sensors – overview and latest technology," 29th August 2017.ttps://www.newark.com/smart-sensors-overview-and-latest-technology (Last accessed January 26, 2019)

[12] DalBello, R. (1985). The land remote sensing commercialization Act of 1984. Space Policy, 1(3), 289-297.

[13] https://docs.microsoft.com/en-us/azure/service-fabric/service-fabric-cloud-services-migration-differences

TOMI: A Framework for Smart Tourism on the Move Innovation

Vaia Moustaka
Department of Informatics
Aristotle University, Thessaloniki, Greece, vmoustag@csd.auth.gr

Athena Vakali
Department of Informatics
Aristotle University, Thessaloniki, Greece, avakali@csd.auth.gr

Nikos Zikos
Movvin PC
Thessaloniki, Greece, nikos@movvin.com

Theoklis Tsirakidis
Movvin PC
Thessaloniki, Greece, theoklis@movvin.com

Leonidas G. Anthopoulos
University of Thessaly
Greece, lanthopo@teilar.gr

ABSTRACT

Smart technologies advancements, emerging markets competition and sustainability needs have radically changed tourism and transport sectors. The key features of this change are the exploitation of evolving Big Data in the business intelligence context, the development of customized services tailored to the needs of consumers with the purpose of improving their experience, and the development of new business models based on the interaction between business and consumers. This is due to the capacity of smart transport technologies to integrate customers sensing and in this way a novel framework aimed at: i) developing personalized transport services in the tourism sector and ii) creating and delivering patterns of tourist consumer behavior according to specific target groups and market segments at tourist destination or country level is designed and outlined. The proposed "TOMI" framework, exploits tour data analytics, in order to enable the deployment of personalized tour services that will be beneficial for tour operators, travellers and any other interested parties (local stakeholders, tourism entrepreneurs, etc.). The exploitation of the "TOMI" framework for the purpose of organizing tours in a city is also addressed through a case study on the city of Thessaloniki.

CCS CONCEPTS

• **Information systems~Information systems applications;** • **Information systems~Data analytics;** • **Applied Computing~Transportation;** • Human-centered computing~Social network analysis

KEYWORDS

tour analytics; road passenger transport services; tourism innovation; smart frameworks; smart tourism; smart transport

ACM Reference format:

Vaia Moustaka, Athena Vakali, Nikos Zikos, Theoklis Tsirakidis and Leonidas G. Anthopoulos. 2019. TOMI: A Framework for Smart Tourism on the Move Innovation. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317051

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317051

1 Introduction

Tourism has widely recognized the need for a more personalized, customer-focused approach recently, which mainly assesses the needs, preferences and requirements of tourists in order to improve travel experience and maximize satisfaction, which are proved to be important [1], [2]. The acquisition of profound knowledge and its proper use, which is achieved through Business Intelligence [3] and Big Data Analytics [4], is a prerequisite for decision making and the design of quality services in the tourism industry [5]. In the time of *Tourism 1.0*, where the presence of Information and Communication Technologies (ICTs) was almost non-existent, the collection of information required long and complex procedures and it was difficult to create precise patterns of tourism profile. Subsequently, the tourism industry, following the rapid ICTs advancements, was transformed from Tourism 1.0 to *Tourism 2.0*, utilizing the capabilities offered by Web 2.0 [6], [7]. As stated by [8] not only the Internet and Web 1.0, but also Web 2.0 and the entire social networks environment had powerful impact on the tourism industry.

Then, the idea of *Tourism 3.0* emerged, which focuses on the promotion of tourist destinations at local and regional level and the active engagement of local actors (i.e., authorities, entrepreneurs, citizens and tourists) in the co-creation of tourism services. In this context, particular emphasis is given to rebranding of local destinations, developing new business

models and tourism activities, storytelling and redefining the role of tour agents [9].

Nowadays, we are living in the *Tourism 4.0* era [10], which describes the trend of the market to take advantage of the content created by users via web or mobile applications with the purposes of creating personalized services and improve travel experience [11]. The main prerequisite for designing and delivering personalized services, which are expected to increase the satisfaction of tourists and lead to lower costs and greater security, is the knowledge acquisition through the exploitation of advanced technologies such as cloud computing, web and mobile applications, virtual reality, Internet of Things, etc. as well as cutting-edge analytics such as artificial intelligence, machine learning algorithms, chat bots, etc. [12–14]. Since the concept of Tourism 4.0 matures, the landscape in travel arrangements that includes accommodation, intermodal transport and tours is changing [15]. A wealth of travel planning and booking options are already available to tourists, who can easily and instantly receive the information they need from various available applications [16]. Moreover, innovative pooled (shared) mobility models that are being developed such as bus pooling, ride pooling, etc., as well as, the efficient transport management through analytics and forecasting methods, lead to reduced costs and resources savings [17], [18]. On the other hand, new challenges are emerging in the tourism market as entrepreneurs will have to adapt to new conditions and redesign their services and marketing policies to meet new requirements, remain competitive and ensure their sustainability. As it is evidenced, the need for travel companies to offer value added services in order to differentiate themselves against the competition is enormous [19].

The shift of the tourism industry on the one hand to the promotion of tourist destinations, introduced by Tourism 3.0, and on the other hand to personalized service co-creation that will improve the tourist experience in the context of Tourism 4.0, are raising challenges for modern cities, which tend to become facilitators of tourism and business activities. Highlighting the strengths and unique characteristics of cities combined with the improvement of tourism services can generate comparative advantages that will distinguish them among other cities in the world's tourism industry. An indicative city example that has exploited these opportunities to become one of the most popular tourist destinations in the world is Barcelona. The capitalization of Olympic Properties and the harbor, which is the fourth cruise port in the world, the provision of high quality public transport, and the preservation of cultural heritage (e.g., Gaudi buildings, Picasso museum, etc.) attract millions of visitors each year [20].

Taking into account that the impact of transport on tourism has been little discussed and that there is a lack of organized and personalized travel services by coaches, inside and outside cities (as discussed later in subsection 2.2), this article aims to propose a novel framework that utilizes tour data analytics and can modernize *Road Passenger Transport Operators (RPTOs)* and support the deployment of customized tour services. The use and benefits of the proposed framework in the tourism sector are addressed in a case study on the city of Thessaloniki.

The contribution of this article is twofold: i) deals with the necessity of data analytics in the field of *RPTOs* with the purpose of acquiring knowledge and designing services for tourists on the move and re-introduces the term *"tour analytics"*, and ii) presents a novel framework along with two new plugins, which were designed in the context of our ongoing project to assist in the modernization of RPTOs and development of customized tour services. The proposed framework, which is extensible, along with the two new add-ons, is expected to directly benefit tour operators and travellers, and indirectly cities and other interested parties.

The rest of this article is organized as follows: *Section 2* deals with tourism analytics, and re-introduces the term *"tour analytics"*, proposing a new definition. The novel proposed *"TOMI" framework* that aims at the modernization of road passenger transport services and the development of personalized tour experiences both inside and outside cities is proposed in *Section 3*. A case study, which examines the exploitation of the proposed framework for the purpose of organizing tours in a city, is presented in *Section 4*. Finally, *Section 5* contains some conclusions and future perspectives.

2 From Tourism to Tour Analytics

The purpose of this section is to clarify the terms: *tourism analytics* and *tour analytics*, and to discuss their contribution to the tourism industry.

2.1 Tourism challenges addressed through Big Data Analytics

The exponential growth of Big Data [21], which may be structured, semi-structured or unstructured heterogeneous data derived from various sources such as online social networks (OSNs), sensors, telecommunication providers, GIS, etc., has significantly affected the tourism industry [5]. The exploitation of this data, which relates to tourists' choices before, during and after the trip, through data mining and forecasting methods for the purpose of gaining unprecedented insights into tourism, is known as t*ourism analytics* [20], [22]. Such analytics constitute an important tool for any tourism and local stakeholder as provides them with the opportunity to better understand the preferences of tourists, to improve their services and to create competitive and attractive tourism packages.

Tourism analytics, which, compared to traditional methods, offer reliability, new data flows and the possibility of using real-time data and forecasting have been used in several studies [22]. The user-generated content (UGC) posted by users on online platforms (e.g., OSNs, TripAdvisor[1] , Booking[2] , etc.) is definitely one of the most useful data sources as it provides direct and unbiased information about the preferences and opinions of tourists regarding to tourism services [23], [24]. Of particular interest is the work of [25] which focused on data analysis to gain knowledge about *tourist destinations*. The development of the "Destination Management Information System Åre" (DMIS

[1] https://www.tripadvisor.com/
[2] https://www.booking.com/index.el.html

Åre), designed by them, offers interested parties (i.e., tour operators, hoteliers, etc.) the ability to know the behavior and experience of the traveler before and after the trip as it can use platforms like Booking.com. Moreover, Brandt et al. [26] investigated the value of spatial and semantic analysis of OSNs messages in *smart tourism ecosystems*, which is a combination of digital ecosystems and smart tourism networks [27], utilizing 600,000 Twitter messages in San Francisco. Their findings have shown that OSNs analysis leads to the extraction of urban spatial patterns related to the presence, environmental engagement and local involvement of users throughout the city, which are useful for the development of *urban smart tourism*. In addition, Vecchio et al. [28], focusing on *smart tourist destinations*, studied a set of regional tourist experiences for the southern European region, and specifically Southern Italy (Apulia region), in order to derive patterns and value creation opportunities from Big Data. As it was demonstrated by their findings, Big Data generated by UGC on OSNs and its real-time integration and analysis is the most important guide to the value creation process in a *smart tourist destination*. Its exploitation improves decision-making processes, creates marketing strategies with personalized offers, contributes to transparency and confidence in dialogue with customers and stakeholders, and leads to the emergence of new business models. Finally, Marine-Roig & Clavé [20] also focused on *smart tourist destinations*, utilizing UGC in travel blogs and websites that included tourists' reviews over the last decade concerning the city of Barcelona.

Apart from the data collected from OSNs and the Internet, data from the telecommunications network utilized in the context of the tourism industry. Leng et al. [29] used call detail records (CDR) as part of the analysis and assessment of tourism strategies at local and national level in the country of Andorra. Furthermore, some scholars combined data from different sources. For instance, Li et al. [30], in order to predict the *travel destination demand* including the *number of tourists* and *hotel occupancy* in Beijing, combined data from Baidu[3] and the Beijing Tourism Association[4].

2.2 The necessity of data analysis in the field of Road Passenger Transport

The planning of *OD-transport (origin-destination transport)*, which describes transport from the tourist home to the tourism destination, as well as, of *local transport*, which is the transport at destinations, is particularly important in the tourism industry [31]. The combination of transport modes in the tourism sector leads to the development of both *international* and *local transport networks* creating new opportunities and markets for tourism destinations [31].

With regard to local transport, tourists choose the means of transport taking into account the following factors: i) time limit, ii) distance, iii) status, iv) comfort, v) security, vi) benefit, vii) price, viii) geographical position, and ix) competition [32]. Taxis and cars are expensive and affect negatively the environment regarding energy use, carbon dioxide emissions, poor air quality, and noise [33], [34]. On the other hand, Guiver et al. [35] in their review, investigating the role of buses in environmental, social and economic sustainability in the tourism sector, found out that buses achieve moderate displacement from cars, allow access to the countryside for people without cars and strengthen local economies. Due to these advantages, they suggested the promotion of buses' use and market segmentation to the identification of new target groups. In addition, urban public transport compared to taxis and rental cars offers low cost and low environmental burden. Nevertheless, public transport, due to supply constraints, confronts passengers' congestion and mitigates the travel experience of tourists by returning them to the daily routine [36].

In this respect, the use of private hire buses (or coaches) of multiple sizes and seating capacity is an ideal solution for tourists' transport, as they gather the benefits of public transport and can provide personalized services, improving travel experience [37]. This service falls within the *road passenger transport mode*, which refers to the total movement of passengers using inland transport on a given network [38]. According to Regulation (EC) No 1073/2009[5] , which governs the national and international carriage of passengers by bus or coach, *RPTOs* are authorized to offer road transport services in various sectors, such as tourism, school transport, personnel transport, etc. after obtaining a Community License.

With regard to tourism, RPTOs can offer personalized bus tours to tourists, turning cities into tourism destinations and improving their experience. Lumsdon [39], investigating the factors that influence the development of tourism bus services, suggested a model for the design of the tourism bus networks. According to this model, some of the main factors to be taken into account for the prevalence of tour buses are quality and service delivery, travel integration between transport and tourism operators, driver behavior, marketing and monitoring mechanisms. Consequently, the design of efficient and competitive services requires knowledge that will come from the appropriate data analytics which, to the best of our knowledge are still missing for services provided by RPTOs [39]. Adopting and revising the term of tour analysis used by [40], who used data from the Sydney Household Travel Survey to compare travel analysis and tour analysis in Sydney, we propose the following definition for tour analytics:

"Tour analytics is data analytics for acquiring profound knowledge and designing transport services for various purposes (e.g., pleasure, business, education, etc.)."

3 Data Aware and Adaptive RPTOs

This section discusses the need for modernization of RPTOs, as well as, the "TOMI" framework, which was designed to address this need.

[3] http://ir.baidu.com/phoenix.zhtml?c=188488&p=irol-irhome
[4] http://www.wta-web.org/eng/hymd_4014/yz/bjsdlyjtyxzrgs/
[5] https://ec.europa.eu/transport/modes/road/passenger-transport_el

3.1 Technology transfer for adaptive RPTOs

The purpose of this article is to present the main idea of our ongoing project that aims at the modernization of RPTOs and the development of customized tourism transport services, utilizing tour analytics. The idea is inspired from the Airline industry, where advanced technologies are used for service providing, demand forecasting, and the development of competitive sales policy [41]. RPTOs in most places around the world are in the same state as airlines were in the 70s [42], [43]. Since then, airlines met a tremendous growth in the technological field, in the management efficiency, in scalability and in the field of profit optimization. On the other hand RPTOs have been left behind and in the time that this article is written most of them are even online unreachable [43]. This lack motivated us to adopt and adapt the airline established technologies in road transport developing a novel framework, which is based on an existing platform entitled *"TOMI" platform*, with the purpose of helping RPTOs to increase their product quality, competitiveness, income and to reduce their operating costs. In addition, this framework is expected to offer personalized services to travellers who use it for ticket reservations as well as valuable knowledge to local authorities who want to make use of local transport networks for various purposes and to any other interested party.

3.2 The "TOMI" framework

The proposed framework that aspires to implement the above idea, enriching the services of the existing *"TOMI" platform* with the addition of two new plugins, is presented in Fig. 1.

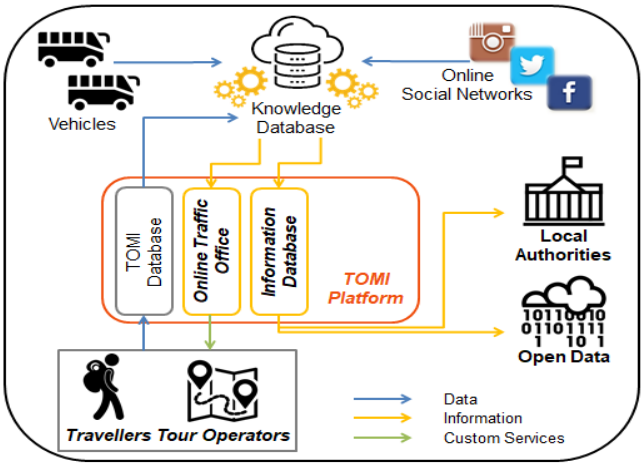


Figure 1: The "TOMI" framework

The *"TOMI" proposed platform* was designed on the basis of the already operational private startup solution *Movvin*[6], which develops integrated solutions for RPTOs. The "TOMI" framework aims to act as a one-stop-shop for RPTOs in the field of management bookings, tickets, vehicles, drivers, routes, etc., all of them integrated in a single online and extensible platform. The operator can handle all online and offline bookings in an efficient way, while online booking are handled automatically with plugins that "TOMI" offers. Those plugins can be plugged in static web pages and transform them into online booking endpoint for travellers. In addition, the platform offers to the traffic officer the ability to manage routes and itineraries in a more efficient way in comparison with traditional management tools In addition, the platform offers to the traffic officer the ability to manage routes and itineraries in a more efficient way in comparison with traditional management tool as AI patterns would be used to undertake day-to-day operational functions (optimal route calculation, route design, etc.). Besides that, the platform offers *Rostering Management tools* [44], which are very useful for large operators or for those who want to increase their scalability. At last the platform offers an enhanced *Fleet Management too*[7]*l* for monitoring vehicles and informs the traffic officer, in real time, for its fleet status.

As demonstrated in Fig. 1, heterogeneous Big Data derived from three different data sources will be exploited for the development of the new plugins. These data sources are as follows:

- ***"TOMI" database:*** The "TOMI" platform handles all bookings, all routes, all itineraries etc., from all RPTOs. When a booking is done, "TOMI" tracks down all related metadata to the specific booking for post process analysis and research. Also, RPTOs handle all routes and itineraries through the platform and this enables "TOMI" to have a clearer view of the transportation as a product in all its aspects and in a wider geographical field. In addition, Rostering Management related data is handled and kept in the database for a better and efficient management in wide time periods using cutting-edge Data Analytics technologies.
- ***FCD and IoT on vehicles:*** Vehicles from the collaborating RPTOs are monitored using IoT devices plugged into the vehicle's ECU[8] (Engine Control Unit). This IoT device is connected using OBD2[9] protocol and performs read only operations. It reads all available data from the *vehicle's status* including engine status, engine failures, other vehicles failures, fuel consumption etc. Moreover, the IoT device has auxiliary systems that read GPS signal, inertial movements or specific emission analyzer for real time gas monitoring. This IoT device is equipped with a cellular modem and through standard mobile protocols, such as GPRS, 3G, 4G, it transmits the data into *Knowledge Database*. Each IoT device performs real time streaming of all those data, which are analyzed and broadcasted to the corresponding operator.
- ***OSNs:*** As pointed out by [45], OSNs (i.e., Twitter, Facebook, Instagram, etc.) act as "human sensors", which record human activity, offering volumes of heterogeneous data at a very low cost. The content created and shared by users in these networks is an important data source about their

[6] https://Movvin.com/

[7] https://whatis.techtarget.com/definition/fleet-management
[8] https://www.computerhope.com/jargon/e/ecu.htm
[9] https://www.csselectronics.com/screen/page/simple-intro-obd2-explained/language/en

preferences, opinions and reviews, which can be exploited in many fields, such as marketing, psychology, smart cities, etc. [46], [47]. In the context of the "TOMI" framework, the data coming from OSNs will be used to gain insights into travellers' preferences and satisfaction regarding tours and RPTOs' services.

This data will be transferred into *Knowledge Database* where will be aggregated and tour analytics will be carried out leading to patterns and insights into itineraries, travellers preferences, vehicles and drivers behavior, etc. Subsequently, this valuable information produced will be transferred into the "TOMI" platform, feeding the two plugins. Finally, the *Online Traffic Office* will provide the traffic offices of RPTOs and travellers with customized services, while the *Information Database* will provide *Local Authorities* (e.g., municipalities, etc.) and *Open Data Platforms* with valuable information.

The "TOMI" framework is expected to have important impact. With regard to RPTOs (tour operators), *fuel consumption optimization* will be performed by suggesting more efficient routes or drive patterns, while in the field of *booking*, optimization in the revenues using will be performed using adaptive fare structure, a well-established technique in the Airlines. Furthermore, last Rostering and Fleet management will be aided by smart optimizers which will help the Traffic Officers to perform scheduling in large scale efficiently. As far as travellers are concerned, personalized services will be developed, while competition among RPTOs will lead to more qualitative and economical tour packages. Finally, local authorities and any other interested parties can make use of the available information, offered by "TOMI", in decision-making and in the design of new services.

4 Proof of Concept Design and Plan

Since the "TOMI" framework is in its development phase, a case study is presented, which explores how a city can become a tourist destination through its use.

Situated in northern Greece, Thessaloniki is the second largest Greek city, which hosts the second largest export and transit port and the second largest international airport of the country. Built near the sea, it is a modern metropolis with rich history, cultural and spiritual heritage and cosmopolitan character, which give it a unique beauty and charm. Thanks to its unique history and interconnectivity, and serving as a base for nearby destinations, Thessaloniki attracts millions of visitors, annually [48]. However, despite its cultural wealth and its great potential to become a top tourism destination, such as Barcelona [20], Thessaloniki lags behind in the field of transport and organized tours. The supply of public urban transport for tourism purposes is limited, with the result that the tourist product and the travel experience of the visitors are degraded.

In this respect, the "TOMI" platform combined with the proposed framework seems to be excellent tools for organizing value added tours and improving the quality of tourist services offered in Thessaloniki. Tour operators (i.e., RPTOs) and travellers are expected to benefit from the customized services offered by the Online Traffic Office, as shown in Fig. 1. In particular, travellers' preferences and reviews, the popularity of attractions and points of interest, and information on the fleet state (e.g., fuel consumption, travelled distance, maintenance, etc.) allow RPTOs to organize attractive and profitable packages, adjust demand-based service prices and manage their resources effectively. With regard to travellers, they will benefit from competitive prices and the quality of services tailored to their needs and preferences. Finally, the city will benefit as its sights and culture will be promoted attracting new visitors, new business activities will be created and competitive advantages will be acquired, turning it into a popular tourist destination.

5 Conclusions

The article proposes a novel framework, called "TOMI", aiming at bridging the gap in supply of customized transport services by buses in the tourism sector. The proposed framework attempts to modernize RPTOs and lead to the deployment of personalized tour services that will be beneficial for tour operators, travellers and any other interested parties (local actors, tourism entrepreneurs, etc.). Exploiting tour data analytics, the "TOMI" framework aspires to expand the services and impact of the "TOMI" platform on which it is based, with two new plugins which are the following: i) *Online Traffic Office* that will offer customized services to RPTOs and travellers, and ii) *Information Database* that will provide valuable insights into urban tourism to local authorities through an open data platform. These plugins is expected to make an effective contribution to fuel consumption optimization, to travellers' preferences identification and to emergence of new tourist destinations. In addition, in the field of booking, revenue growth is expected through the use of an adaptive structure of fares, while travellers will benefit from competition among tour operators. The use and benefits of the "TOMI" framework in the tourism industry are investigated in a case study on the city of Thessaloniki.

Our future thoughts concentrate mainly on three directions. The first one involves the implementation of the "TOMI" framework through the development of the proposed plugins to be integrated into the "TOMI" platform and its extension through market research and design of new services directly or indirectly related to road passenger transport. The second concerns the thorough discussion of the smart mobility and sharing issues, and the comparison of the proposed framework with the car pooling services. Finally, the latter aims at analyzing the new business models in road transport and the needs that led to their development, as well as assessing the impact of the "TOMI" framework on the value chain of a city and on the operation of RPTOs.

6 ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their valuable suggestions which have contributed in the paper's better focus and readability. This research has been co - financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and

Innovation, under the call RESEARCH – CREATE – INNOVATE (Project Code: T1EDK-03052).

REFERENCES

[1] A., Correia, M., Kozak, and J. Ferradeira (2013). From tourist motivations to tourist satisfaction. International Journal of Culture, Tourism and Hospitality Research , 7(4), 411–424.

[2] M., Choudhury, and P., Harrigan (2014). CRM to social CRM: The integration of new technologies into customer relationship management. Journal of Strategic Marketing, 22(2), 149–17.

[3] R., Baggio (2016). Big Data, Business Intelligence and Tourism: a brief analysis of the literature. Available at: https://www.iby.it/turismo/articles/baggio_BigDataSurvey.pdf (Accessed Jan. 2019)

[4] A.,Gandomi, and M., Haider (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35, 137–144. DOI: http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007

[5] Gartner (2018). Business Intelligence. Available at: https://www.gartner.com/it-glossary/business-intelligence-bi/ (Accessed Jan. 2019)

[6] R. van Hooijdonk (2015). Technology Trends 2030. Guest lecture retrieved on December 10th, 2015 from: NHTV University of Applied Sciences, Breda, The Netherlands.

[7] M., Pilkington (2017). Tourism for Development in the Republic of Moldova: Empowering Individuals and Extending the Reach of Globalization. Handbook of Research on Individualism and Identity in the Globalized Digital Age. DOI: 10.4018/978-1-5225-0522-8.ch022 (Accessed Jan. 2019)

[8] R., Law, D., Buhalis, and C., Cobanoglu (2014). Progress on information and communication technologies in hospitality and tourism. International Journal of Contemporary Hospitality Management, 26(5), 727-750.

[9] J.P., Segarra (2015). The Vision of Tourism 3.0 (White article). Available at: https://www.slideshare.net/JordiPera/the-vision-of-tourism-30-whitearticle (Accessed Jan. 2019)

[10] Tourism 4.0. (2018). What is Tourism 4.0?. Available at: https://www.tourism4-0.org/ (Accessed Jan. 2019)

[11] E., Oliveira, and E., Panyik (2014). Content, context and co-creation. Journal of Vacation Marketing, 21(1), 53–74. DOI:10.1177/1356766714544235

[12] A., Nabben, E., Wetzel, E., Oldani, J., Huyeng, M., Van de Boel, Z., Fan, C., Barten, C. (2016). Smart Technologies in Tourism: Case study on the influence of iBeacons on customer experience during the 2015 SAIL Amsterdam event. In International Tourism Student Conference, 19-22 April 2016, Madrid, Spain.

[13] G., Stalidis, Dimitrios Karapistolis, Athanasios Vafeiadis (2015). Marketing Decision Support Using Artificial Intelligence and Knowledge Modeling: Application to Tourist Destination Management. Procedia - Social and Behavioral Sciences, 175(12), Feb. 2015, 106-113. DOI: https://doi.org/10.1016/j.sbspro.2015.01.1180

[14] D., Newman (2018). Top 6 Digital Transformation Trends in Hospitality and Tourism. Forbes. Available at: https://www.forbes.com/sites/danielnewman/2018/01/02/top-6-digital-transformation-trends-in-hospitality-and-tourism/ (Accessed Jan. 2019)

[15] J. E., Dickinson, K., Ghali, T., Cherrett, C., Speed, N., Davies, S., Norgate (2012). Tourism and the smartphone app: capabilities, emerging practice and scope in the travel domain. Current Issues in Tourism 17, 1, 84–101.doi:10.1080/13683500.2012.718323

[16] TravelCarma (2018). How Mobile is Reshaping the Travel & Tourism Industry. Available at: https://blog.travelcarma.com/2017/08/29/mobile-reshaping-travel-tourism-industry/

[17] V. Crabtree (2007). Public Transport is Good. The Advantages of Pooled Transport. Available at: http://www.vexen.co.uk/notes/public_transport.html (Accessed Jan. 2019)

[18] S., Ram, Y., Wang, F., Currim,F. Dong, E. Dantas, L.A., Sabóia (2016). SMARTBUS: A Web Application for Smart Urban Mobility and Transportation. In Companion Proceedings of the 25th International Conference Companion on World Wide Web, 363-368. DOI: https://doi.org/10.1145/2872518.2888613

[19] WYRED (2018). Effects of Industry 4.0 on Tourism Industry – Development from E-Tourism to M-Tourism. Available at: https://wyredproject.eu/2018/04/17/effects-of-industry-4-0-on-tourism-industry-development-from-e-tourism-to-m-tourism/ (Accessed Jan. 2019)

[20] E., Marine-Roig, and S., Clavé (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. Journal of Destination Marketing & Management. DOI: http://dx.doi.org/10.1016/j.jdmm.2015.06.004

[21] Bernard Marr (2015). Why only one of the 5 Vs of Big Data really matters. (March 2015). Available at: http://www.ibmbigdatahub.com/blog/why–only–one–5–vs–big–data–really–matters (Accessed Jan. 2019)

[22] Haiyan Song, and Liu Han (2017). Predicting Tourist Demand Using Big Data. Analytics in Smart Tourism Design. DOI: 10.1007/978-3-319-44263-1_2

[23] C., Koltringer, and A., Dickinger (2015). Analyzing destination branding and image from online sources: A web content mining approach. Journal of Business Research, 68(9), 1836–1843. http://dx.doi.org/10.1016/j.jbusres.2015.01.011

[24] W., Lu, and S., Stepchenkova (2015). User-generatedcontentasaresearchmodein tourism and hospitality applications: Topics, methods, and software. Journal of Hospitality Marketing & Management, 24(2), 119–154. DOI: http://dx.doi.org/10.1080/19368623.2014.907758

[25] M., Fuchs, W., Höpken, and M., Lexhagen (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. Journal of Destination Marketing & Management (2014), DOI: http://dx.doi.org/10.1016/j.jdmm.2014.08.002i

[26] T., Brandt, J., Bendler, and D., Neumann (2017). Social Media Analytics and Value Creation in Urban Smart Tourism Ecosystems, Information and Management. DOI: http://dx.doi.org/10.1016/j.im.2017.01.004

[27] U., Gretzel, H., Werthner, C., Koo, and C., Lamsfus (2015). Conceptual foundations for understanding smart tourism ecosystems, Computers in Human Behavior, 50, 558–563. DOI: https://doi.org/10.1016/j.chb.2015.03.043

[28] P., Vecchio, G., Mele, V., Ndou, G., Secundo (2018). Creating value from Social Big Data: Implications for Smart Tourism Destinations, Information Processing and Management 54 (2018), 847–860. DOI: http://dx.doi.org/10.1016/j.ipm.2017.10.006

[29] Y., Leng, A., Noriega, A.S, Pentland, I., Winder, N., Lutz, and L., Alonso (2016). Analysis of Tourism Dynamics and Special Events through Mobile Phone Metadata. In Proceedings of Data for Good Exchange (D4GX) 2016. New York, NY.

[30] X., Li, B., Pan, R., Law, X., Huang (2017). Forecasting tourism demand with composite search index. Tourism Management, 59. DOI: http://dx.doi.org/10.1016/j.tourman.2016.07.005

[31] G., Lohmann, and Duval, D., T. (2011). Critical Aspects of the Tourism-Transport Relationship. Contemporary Tourism Review. C. Cooper. Oxford, Goodfellow Publishers.

[32] J., Westlak, and D. Robbins (2005). Transportation in Cooper C., Fletcher J., Fyall A., Gilbert D. and S. Wanhill (eds), Tourism: Principles and Practice, 3rd edition, Pearson Edu¬cation Limited, Essex.

[33] S., Becken, D. Simmons, and C. Frampton (2003). Energy Use Associated with Different Travel Choices. Tourism Management, 24, 267–277.

[34] C., Speakman (2005). Tourism and transport: Future prospects, Tourism and Hospitality Planning & Development, 2(2), 129-135, DOI: 10.1080/14790530500173647

[35] J., Guiver, L., Lumsdon, R., Weston & M., Ferguson (2007). Do buses help meet tourism objectives? The contribution and potential of scheduled buses in rural destination areas. Transport Policy, 14(4), 275–282. DOI:10.1016/j.tranpol.2007.02.006

[36] D., Albalate, and G., Bel (2010). Tourism and urban public transport: Holding demand pressure under supply constraints. Tourism Management, 31(3), 425–433. DOI:10.1016/j.tourman.2009.04.011

[37] R., Carreira, L., Patrício, R., Natal Jorge, C., Magee, and Q., Van Eikema Hommes (2013). Towards a holistic approach to the travel experience: A qualitative study of bus transportation. Transport Policy 25, 233–243. DOI:10.1016/j.tranpol.2012.11.009

[38] OECD (2019). Passenger Transport. Available at: https://data.oecd.org/transport/passenger-transport.htm (Accessed Jan. 2019)

[39] L.M., Lumsdon (2006). Factors affecting the design of tourism bus services. Annals of Tourism Research, 33(3), 748–766. DOI:10.1016/j.annals.2006.03.019

[40] F., Milthorpe, and A. Daly (2010). Comparison of Trip and Tour Analysis of Sydney Household Travel Survey Data. Australasian Transport Research Forum 2010 Proceedings, 29 Sep. – 1 Oct. 2010, Canberra, Australia. Available at: http://www.patrec.org/atrf.aspx (Accessed Jan. 2019)

[41] P., Belobaba, A., Odoni, and C., Barnhart. 2015. The global airline industry. 2nd Ed., Wiley.

[42] K. M. Gwilliam (2002). Cities on the Move: A World Bank Urban Transport Strategy Review. Worldbank, 2002. Available at: http://siteresources.worldbank.org/INTURBANTRANSPORT/Resources/chapter7.pdf (Accessed Jan. 2019)

[43] OECD. 2009. The future for interurban passenger transport. Bringing citizens closer together. In 18th International Symposium on Transport Economics and Policy, 16-18 November 2009, Madrid. Available at: https://www.itf-oecd.org/sites/default/files/docs/10interurban.pdf (Accessed Jan. 2019)

[44] A., Ernst, H., Jiang, M., Krishnamoorthy and D., Sier (2004). Staff scheduling and rostering: A review of applications, methods and models. European Journal of Operational Research, 153(1), 3–27. DOI:10.1016/s0377-2217(03)00095-x

[45] D. Doran, K. Severin, S. Gokhale, and A. Dagnino (2014). Social Media Enabled Human Sensing for Smart Cities. AI Communications 2014. Retrieved, Dec. 2017, from: http://knoesis.wright.edu/sites/default/files/aic14.pdf

[46] V. Moustaka, A., Vakali, L.-G., and Anthopoulos (2019). A Systematic Review for Smart City Data Analytics. ACM Comput. Surv., 51(5), Article 103 (Jan. 2019), 41 pages. DOI: https://doi.org/10.1145/3239566

[47] V., Moustaka, Z., Theodosiou, A., Vakali, A., Kounoudes, L.-G., Anthopoulos (2018). Enhancing Social Networking in Smart Cities: Privacy and Security Borderlines. Technological Forecasting & Social Change, Elsevier. DOI: https://doi.org/10.1016/j.techfore.2018.10.026

[48] GTP-Headlines (2018). More Foreign Visitors at Thessaloniki Hotels in 2017. Available at: https://news.gtp.gr/2018/02/23/more-foreign-visitors-thessaloniki-hotels-2017/ (Accessed Jan. 2019)

Travel Time Estimation Accuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR*

Daksh Shah
IIIT-Delhi
daksh17336@iiitd.ac.in

Aravinda Kumaran
IIIT-Delhi
aravinda15133@iiitd.ac.in

Rijurekha Sen
IIT-Delhi
riju@cse.iitd.ac.in

Ponnurangam Kumaraguru†
IIIT-Delhi
pk@iiitd.ac.in

ABSTRACT

Travel time estimates are highly useful in planning urban mobility events. This paper investigates the quality of travel time estimates in the Indian capital city of Delhi and the National Capital Region (NCR). Using Uber mobile and web applications, we collect data about 610 trips from 34 Uber users. We empirically show the unpredictability of travel time estimates for Uber cabs. We also discuss the adverse effects of such unpredictability on passengers waiting for the cabs, leading to a whopping 28.4% of the requested trips being cancelled. Our empirical observations differ significantly from the high accuracies reported in travel time estimation literature. These pessimistic results will hopefully trigger useful investigations in future on why the travel time estimates are mismatching the high accuracy levels reported in literature - (a) is it a lack of training data issue for developing countries or (b) an algorithmic shortcoming that cannot capture the (lack of) historical patterns in developing region travel times or (c) a conscious policy decision by Uber platform or Uber drivers, to mismatch the correctly predicted travel time estimates and increase cab cancellation fees? In the context of smartphone apps extensively generating and utilizing travel time information for urban commute, this paper identifies and discusses the important problem of travel time estimation inaccuracies in developing countries.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

ACM proceedings; LaTeX; text tagging; use semi-colons

ACM Reference Format:
Daksh Shah, Aravinda Kumaran, Rijurekha Sen, and Ponnurangam Kumaraguru. 2019. Travel Time Estimation Accuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317057

*Produces the Permission Block, and Copyright Information
†This work was done in part while Ponnurangam Kumaraguru was on sabbatical at IIIT Hyderabad.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317057

1 INTRODUCTION

Travel time estimates are extremely important for urban commuters, for appropriate trip planning. Even in developing countries where transport infrastructure growth is slow, travel time estimates are regularly used for different transport services. In India, for example, public transport fleets are getting fitted with GPS tracking devices and their real time trip data are becoming publicly available [2]. There are also proprietary travel time datasets owned by Google, Uber and similar cab sharing services, crowd-sourced from Google Map users or cab passengers and drivers. All these datasets make travel time estimation possible, based on historical and recent trend analyses using a myriad of prediction algorithms. The estimates are in turn consumed by different transport services: e.g. to estimate the arrival time of buses in bus-stops or cabs at the pick-up locations, or to compute the overall trip times in ongoing cab rides or in Google Maps directions app.

This paper investigates the quality of these travel time estimates in the Indian capital city of Delhi and the National Capital Region (NCR). We use Uber's Estimated Time of Arrival (ETA) of cabs, as a proxy for travel time estimates. In India, the ride-sharing service Uber recorded 1 million trips per day in 2017, including a Delhi commuter taking as many as 5 rides per day [1]. The rate of ridership growth slowed down in 2018, but still there was a significant number of 3.5 million rides per day [7]. Delhi contributes about 10-12% of the Indian trips (about a million trips per week) [6].

Uber's travel time estimation algorithms can therefore use the extensive GPS data that Uber collects from its own cabs. The predictions, in turn, affect the routing and scheduling behaviors of its large and rapidly growing community of passengers and drivers. Thus Uber's travel time estimation numbers form an interesting dataset to empirically examine the travel time prediction quality, as inputs to the estimation algorithms are rich and the outputs are in massive use. Uber acknowledges [9] that its "ETA times are estimates and not guaranteed. A variety of external factors like heavy traffic or road construction can impact travel time." In developing country cities like Delhi-NCR, where traffic congestion and construction are rampant [3–5, 8], this paper explores the quality of Uber's ETA using empirical data from real trips.

State of the art research literature in travel time estimation [13, 14] report around 2 minutes of travel time estimation errors. We, however, find 5-10 minutes of median to 80^{th} percentile errors, with 25 minutes errors in the worst case. These values, empirically computed using 610 trips' data from 34 Uber users in Delhi-NCR, give rise to interesting discussion points in this paper. The errors are not coming from manual reporting errors, as we collect information about users' trips using automated crawling of the Uber smartphone and web apps. Whether the errors come from (i) data or algorithmic shortcomings, or (ii) due to conscious decisions taken by the Uber platform or the Uber drivers, are open questions formulated in this discussion paper, to be explored in future.

2 UBER'S ARRIVAL TIME ESTIMATES

While assessing the quality of travel time estimates, we take Uber's Estimated Time of Arrival (ETA) for its cabs, as a proxy for travel time estimates. Uber has a web API, which takes as input a GPS coordinate value and returns the Estimated Time of Arrival (ETA) of a cab at that GPS location. This ETA value returned by the API is the same value as shown in the Uber smartphone app (manually verified by us). The Uber user sees this ETA as soon as he opens the app and his location is detected. We refer this ETA as returned by the API and seen at Uber app start time as *t1*.

However, we observe an interesting phenomenon when a booking request is made by the Uber app user. The displayed ETA changes in the Uber app, sometimes showing a significantly different value from *t1*. Our intuitive understanding of this phenomenon is the ETA shown at app start and returned by the API is based on all available cabs in the passenger's neighborhood when the app is started, whereas the value shown after making a booking is based on the actual driver assigned to the request. We denote this ETA shown after a booking request is made as *t2_first*.

t2_first is called so because as the passenger is waiting for the cab, a series of ETA values are displayed in the app, until the cab finally arrives or the trip is cancelled (either by the driver or by the passenger). This range of displayed ETA values as the passenger is waiting is referred to as *t2* and the first ETA as shown in the app is therefore *t2_first*.

Curiously, *t2* remains constant at the same minute over several minutes and also sometimes jumps up instead of monotonically going down, as Uber recalculates the estimations periodically. We refer to time windows during which *t2* remains constant over more than one minute as *t2_stationary* and the instances where *t2* jumps up as *t2_jumps*. As an example, Fig. 1 depicts two different trips with the same *t2_first* along the y-axis as 11 minutes, while the x-axis shows the actual waiting times. Ideally, both curves should monotonically decrease by one minute every minute and reach 0 after 11 minutes. However, while one trip denoted in green as ride 2 has *t2* reaching 0 in about 11 minutes (albeit with few *t2_stationary* instances), the other trip denoted in red as ride 1 has *t2* reaching 0 in almost 21 minutes, with many more *t2_stationary* and *t2_jumps* instances.

Thus the actual time a passenger has to wait can have low correlation with the estimated travel time displayed in the app. Such unpredictable wait times, where 11 minutes of estimated travel time of the cab to the pickup location can mean exactly 11 minutes or

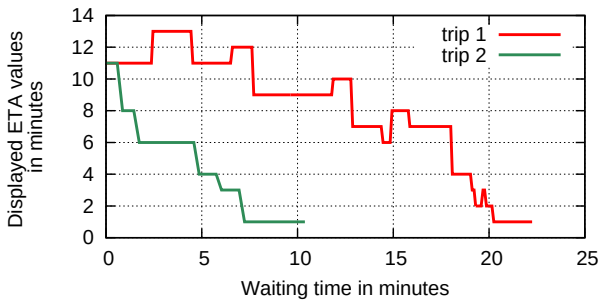


Figure 1: Unpredictable static values and jumps in ETA during waiting, for 11 mins initial ETA

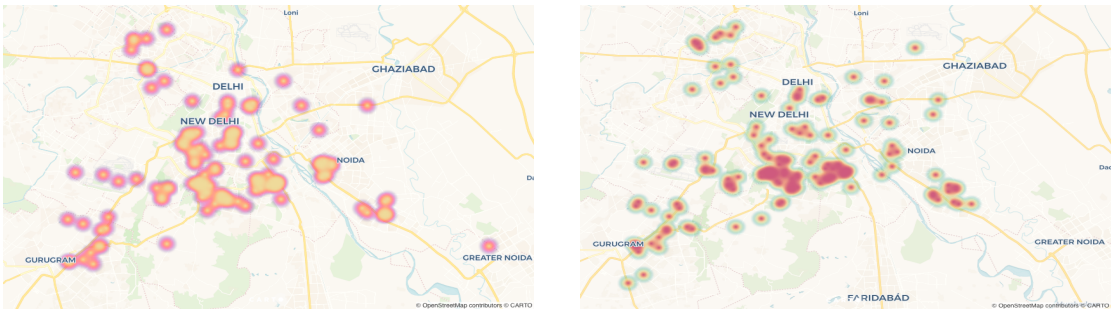
even 21 minutes (91% increase) as seen in Fig. 1, are problematic. It can stress waiting passengers who cannot plan their schedules with confidence, even for important trips. Additionally seeing constant and jumping ETAs displayed in the app while waiting, can increase their frustration. We explore this issue in depth in this paper, by recruiting actual Uber passengers as participants and quantifying the extent to which they experience this unpleasant phenomenon of unpredictable waiting times, due to issues in Uber's travel time estimation.

3 PARTICIPANT RECRUITMENT

Booking Uber cabs would be unnecessary in this study, if we used the Uber web API to get ETA information. But as mentioned above, we have anecdotal evidence that the ETA shown before and after booking (*t1* and *t2_first* respectively) can vary. We will show empirical evidence of this anecdotal observation in Fig. 3. The web API ETA thus loses meaning as soon as an actual booking is made. Also the web API does not record the actual time of a cab's arrival at the pickup point. Due to these two shortcomings of the web API, we need real bookings to compare Uber's travel time estimations with the actual cab arrival times, and quantify errors if any, for meaningful analyses in this paper.

Booking cabs just for the sake of data analyses would violate Uber's term of services. So we recruit real Uber users in Delhi-NCR and collect ride data using their Uber app authorization tokens. Using personal communication channels like email, social media and word of mouth, we advertised our study and received confirmation from 34 Uber users. They agreed to share their ride data, given that all results that we publish are aggregated statistics without revealing their personal information. These users created their own Uber developer apps using their Uber credentials and shared with us the OAuth (authorization) tokens.

Using these authorization tokens we make Uber API calls to detect if the user is currently requesting a cab, waiting for a cab, taking a ride or has experienced a cab cancellation (done by himself or the driver). We collect this data over two months which gives us 610 unique trips, 437 among them being successful and the remaining 173 cancelled. Fig. 2 gives heatmaps of the pickup and destination locations for the recorded rides, showing their significant geographical coverage, though the participant recruitment through our personal contacts might have had some bias.



(a) Pickups for Uber trips

(b) Destinations for Uber trips

Figure 2: Heatmaps of source and destination of the trips made by our recruited participants, showing the significant geographical coverage by the participants

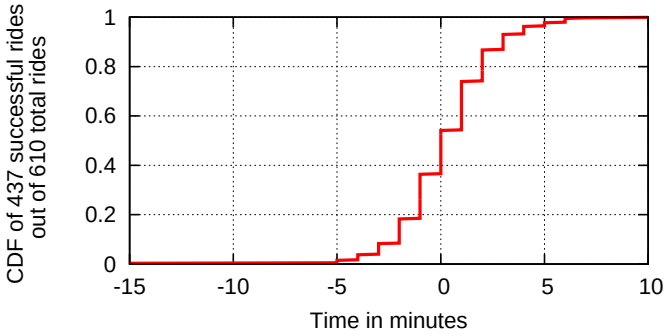


Figure 3: ETA difference before and after booking

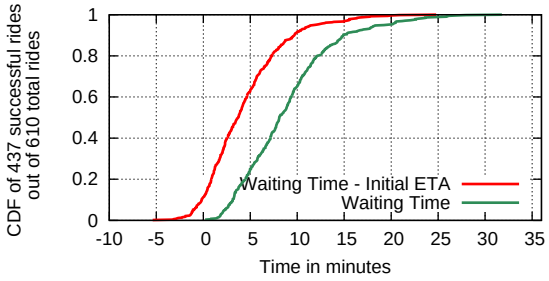


Figure 4: Waiting times and difference between initial ETA displayed after booking and actual waiting times

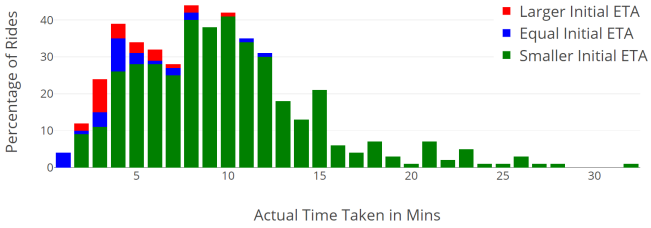


Figure 5: Initial ETA and actual waiting times mismatch

The detailed ETA shown in the app for these 610 trips (*t1*, *t2* with *t2_first* and *t2_last*) and the actual waiting times for the cabs are recorded. The usefulness of participant recruitment is captured in Fig. 3. It plots the CDF of differences between *t1* and *t2_first*, i.e. the ETA displayed before and after booking for the 437 successful trips. This graph computed over the large number of real trips confirm our anecdotal evidence that such difference exist in ETA before and after booking, and therefore motivate our use of real cab booking data to supplement the web API data.

4 PARTICIPANT DATA ANALYSIS

Fig. 4 shows two CDFs, one (in green) depicting the actual waiting time for a cab. It has the top 40^{th} percentile more than 10 minutes for our collected trip dataset. These participants had to wait for more than 10 minutes for their cabs to arrive, which is a significant wait. The median waiting time is 8 minutes.

The second curve in Fig. 4 (in red) shows the CDF of the difference between the ETA shown after booking (*t2_first*) and the actual waiting times before the cab arrives and the trip can start. For less than 20^{th} percentile, the actual waiting times were less than or equal to the ETA, where the curve is to the left of the $x = 0$ line. More common, however, is 5-10 minutes of differences (median to 80^{th} percentile), with the difference going to more than 20-25 minutes in the worst case. Thus the waiting for the cab can be long (more than 10 minutes for top 40^{th} percentile, going upto 35 minutes in the worst case) and there is a gap between the initial ETA shown and the actual waiting time of the user (5-10 minutes differences).

Both these factors of high waiting times and difference between displayed ETA values in the app and the actual waiting times, increase the unreliability of the ride sharing services, as faced by the commuter.

We explore the difference between displayed ETA and actual waiting times in some more detail. Fig. 5 shows the initial ETA after booking along y-xis vs. the actual waiting times along x-axis. It highlights the few instances when the ETA and the waiting times match (in blue), instances where the ETA is an overestimation of the actual waiting times (in red) and instances where the ETA is an underestimation of the actual waiting time (in green). This graph is

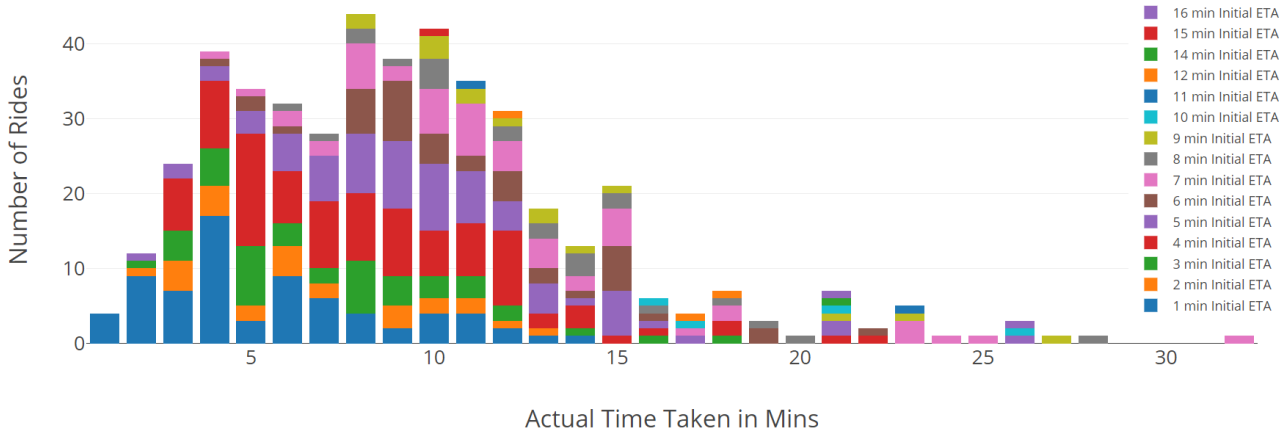
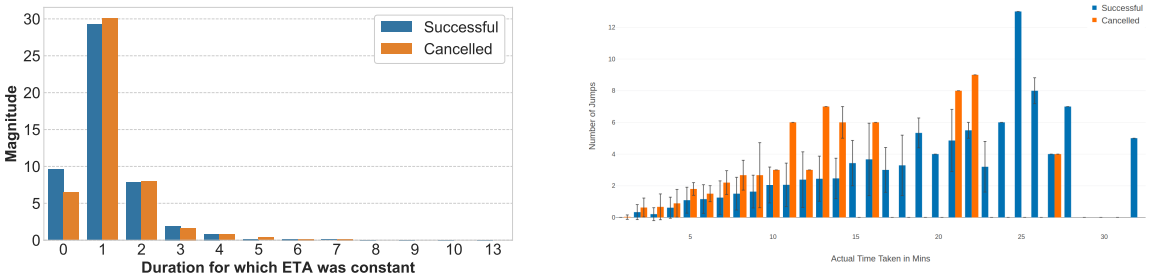


Figure 6: Wide range of initial ETA (different colors represent different initial ETA values) for the same actual waiting time (each bar represents a particular waiting time)



(a) ETA showing constant value for more than a minute

(b) ETA showing upward jumps

Figure 7: ETA remaining constant for more than a minute or jumping to higher values

another depiction of the second CDF in Fig. 4(b). The blue instances with ETA matching the actual waiting times are correct and the red instances overestimating ETA make the users pleasantly surprised when the cab arrives earlier than expected. These are equivalent to the part of the second CDF to the left of the y-axis (x=0) in Fig. 4(b)). The blue and red instances are, however, far less than the green instances of underestimating the actual waiting times, which might make the commuters wary.

Fig. 6 shows a finer granularity of the displayed ETA values along y-axis, with the actual waiting times along x-axis. It further brings out the unreliability of the displayed ETA values. The same actual waiting time can see 9 different initial ETA values (each bar showing upto 9 different colors, each corresponding to a different initial ETA or *t2_first*). The converse way of stating this is the same initial ETA sees a wide range of actual waiting times (each color present in many different bars, each bar corresponding to a different actual waiting time).

Fig. 1 earlier gave an intuition why the same initial ETA (11 minutes in Fig. 1) led to different actual waiting times. During waiting, the ETA remained constant at the same minute value for more than a minute (referred to as *t2_stationary* above) and also jumped up, instead of monotonically decreasing (referred to as *t2_jumps* above). We plot these instances of *t2_stationary* and *t2_jumps* over our 610 collected trips in Fig. 7. The bars for $x >= 2$ in Fig. 7(a) capture *t2_stationary* and Fig. 7(b) shows the average and standard deviation of the number of jumps along y-axis vs. actual waiting times on x-axis.

5 BOOKING CANCELLATIONS

173 out of 610 recorded bookings in our study were cancelled (a whopping 28.4%). Fig. 8 shows the last ETA displayed for successful and cancelled trips. While for successful trips the last ETA is predominantly 0 or 1 minute (the cab arrived after that), for cancelled trips this was higher at 7-9 minutes (the trip was cancelled after that). The fine-grained *t2_constant* and *t2_jumps* information about the cancelled trips, which cause the ETA to remain high during waiting are also shown in Fig. 7. The frustration of the commuters with Uber's travel time estimates is evident from almost 30% of the bookings being cancelled.

Travel time estimates are meant to pacify commuters, keeping them better informed about what to expect. Uber's erratic travel time estimates and their lack of correspondence with actual cab

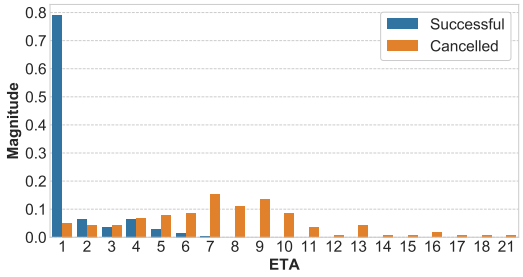


Figure 8: Final ETA for successful vs. cancelled trips

arrival times, would however increase commuter stress. Such stress causes the commuters to cancel the cab booking. Cancellations would need them to try and book another cab from Uber, hoping the new assignment will have a short ETA better matching the actual time of cab arrival, or try a different ride-sharing company like Ola in Delhi-NCR, or try public transport. Cancellations are also not always free [10–12], so in addition to travel inconvenience, the commuters might also incur financial loss.

6 DISCUSSION

Based on our empirical analyses, we now formulate some open questions for future work in this discussion paper.

6.1 Gap between literature and practice

The red line in Fig. 4 shows the median to 80^{th} percentile difference between the actual waiting time for cab arrival and Uber's ETA as 5-10 minutes or 300-600 seconds. The worst case difference is 25 minutes or 1500 seconds. To understand how these numbers compare to the travel time estimation accuracies reported in literature, we refer to two state of the art papers [13, 14].

Table 1: Mean Absolute Error (MAE) of different travel time estimation algorithms as given in [14]. We replicate a subset of Table 2. in [14] here.

Travel time estimation algorithms	Porto MAE (sec)	Shanghai MAE (sec)
RTTE [Rahmani et al., 2013]	169.45	214.01
PTTE [Wang et al., 2014]	159.43	168.48
SVR [Asif et al., 2014]	241.41	424.12
SAE [Lv et al., 2015]	222.06	310.47
spd-LSTM [Ma et al., 2015]	217.37	302.45
TEMP[Wang et al., 2016]	193.61	248.70
Deeptravel[Zhang et al., 2018]	**113.24**	**126.59**

Using Deep Neural Network based algorithms, the authors in [14] report Mean Absolute Error (MAE) of 113.24 seconds for Porto and 126.59 seconds for Shanghai. Using similar algorithms, the authors in [13] report MAE of 186.93 ± 1.01 seconds for Chengdu and 218.29 ± 1.63 seconds for Beijing. In addition to reporting their own MAE values, the authors also report MAE of other travel time estimation algorithms as baselines.

Table 1 and Table 2 give a subset of these values as given in [13, 14]. As can be seen from the rows in bold in the two tables, the MAE values of state of the art algorithms are far less than 300-600 seconds and worst case error of 1500 seconds, observed over 437 successful Uber trips in Figure 4.

Table 2: Mean Absolute Error (MAE) of different travel time estimation algorithms as given in [13]. We replicate a subset of Table 1. in [13] here.

Travel time estimation algorithms	Chengdu MAE (sec)	Beijing MAE (sec)
GBDT	266.15 ± 2.24	393.98 ± 2.99
MlpTTE	265.47 ± 1.53	489.54 ± 1.61
RnnTTE	246.52 ± 1.65	275.07 ± 1.48
DeepTTE	**186.93 ± 1.01**	**218.29 ± 1.63**

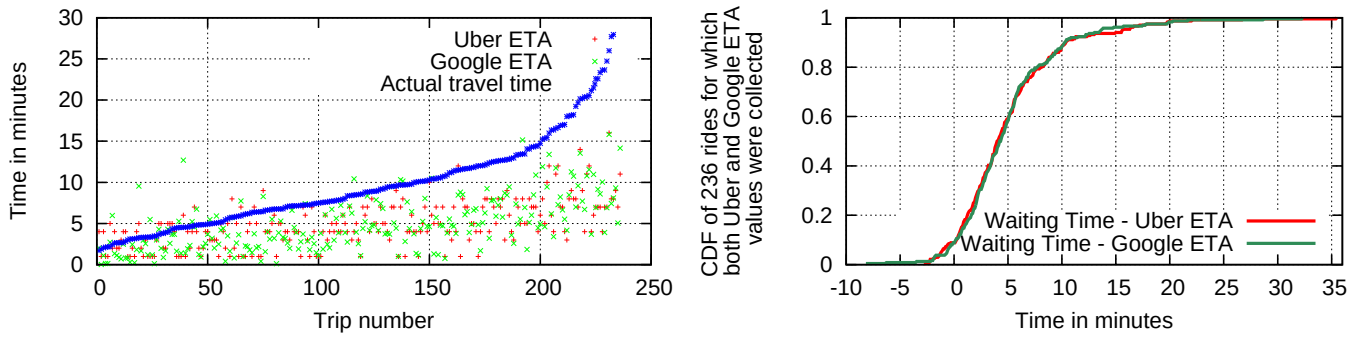
The question is where this gap between Uber's travel time estimates in Delhi-NCR and that in research literature comes from. Is it that Uber does not have enough data to train its algorithms, even though it reports more than a million rides per week in Delhi-NCR [1, 6, 7]? Or is it that the algorithms that Uber uses are unable to capture the (lack of) predictable patterns in Delhi-NCR traffic?

In both Table 1 and Table 2, for the same algorithm in a given row, MAE values differ between two cities of Porto and Shanghai, and that between Chengdu and Beijing. If Delhi MAE were reported in these tables, would the values match our empirical observations of 300-600 seconds MAE? It will be an important direction of future work to examine the limitations of travel time estimation algorithms, in the context of developing countries, to check what MAE the top algorithms give. The challenge will be to get datasets in the scale of Uber or Google, comparable to these urban datasets of Porto, Shanghai, Chengdu and Beijing, to ensure fair comparisons.

6.2 Why is it important to understand the gap

It is important to explore this literature vs. practice gap to check whether the Uber platform intentionally deflates the ETA values or the Uber drivers intentionally mismatch the correctly predicted travel time estimates. These questions arise due to certain characteristic of Uber's cab cancellation policies. In India, if commuters cancel their cab within 5 minutes of booking, they do not incur any cost [10, 11]. However, beyond 5 minutes, the commuters incur a fee. This is the exact opposite cancellation fee policy that Uber has globally [12], where no charges are incurred within 5 minutes of cancellation whereas commuters incur cancellation charges after 5 minutes since booking. This difference in cancellation policies itself is very interesting. What leads to a different policy in India compared to the rest of the world?

In India, therefore, Uber has an incentive to keep passengers waiting for more than 5 minutes, showing them deflated ETA values. The correct high ETA value, if shown to the passenger just after booking, might cause him or her to immediately cancel the booking for free. If Uber intentionally wants to make a profit out of cancellations, then showing deflated ETA values that remains constant or jumps up for more than 5 minutes, is useful. It is also possible that drivers get a share of the cancellation fees, and it is them who intentionally drive slowly or do not start driving at all,



(a) ETA and actual waiting times, for Uber and Google

(b) Difference between ETA and actual waiting times, for Uber and Google

Figure 9: Google vs. Uber ETA values. Though the absolute values are different in (a), the distribution of the differences with actual waiting times is very similar in (b).

to violate Uber's correct time estimates. It is not the Uber platform, but the drivers who want the commuters to cancel the cabs. We need to verify in future Uber's policies for drivers, whether it is possible for them to monitor such intentional delays on the part of drivers and penalize them when detected.

6.3 Preliminary comparisons with Google data

For the last 236 out of the 437 successful rides, we started collecting the Google travel time estimate from the Uber cab's current location to the cab's pickup location, using the Google Directions API. This gave us the Google ETA, corresponding to the Uber ETA, for these 236 trips. Fig. 9(a) shows the absolute values of the actual waiting times and Uber and Google ETA. Since the Google and Uber ETA values do not completely overlap, this indicates that Uber does not call Google's directions API for ETA. It shows its own computed ETA values.

We compute the differences between both these ETA values and the actual waiting times. Fig. 9(b) shows two CDFs of the differences, which look very similar for Google and Uber. Thus though the two companies compute different ETA values, they both differ similarly with the actual waiting times - 20^{th} percentile overestimation where the actual waiting time is less than or equal to the ETA, median to 80^{th} percentile underestimation of 5-10 minutes and 25 minutes underestimation in the worst case.

Our preliminary investigations indicate that this difference between ETA and actual travel times is probably not intentional on part of the Uber platform. Since the two companies independently give similar ETA difference distributions, then either it is actually hard to estimate the ETA correctly (Google and Uber do not have enough data or the correct algorithms), or they actually are giving a correct estimate which Uber drivers are violating to cause cab cancellations. Unless the limitations of data and algorithms are understood with confidence in this context, it will be hard to do these fine-grained transparency studies, involving multiple stake-holders like commuters, drivers and the ride sharing platform. Any claim against a particular stake-holder will be a serious allegation without proof, which should be avoided through data driven understanding of travel time estimation limits in developing regions.

7 CONCLUSION

This paper identifies an important literature vs. practice gap in travel time estimation accuracy in developing regions using empirical data. In future, we will work on quantifying training data and algorithmic limits for this problem in developing regions, by generating large scale travel time datasets. It is necessary to bring more transparency to complex urban mobility services like Uber, and this discussion paper establishes this necessity.

REFERENCES

[1] Indian Express. 2017. Uber recorded 1 million rides a day in India during 2017, Delhi rider took 5 trips daily. Retrieved Mar 7, 2019 from https://indianexpress.com/article/technology/tech-news-technology/uber-recorded-1-million-rides-a-day-in-india-in-2017-\delhi-rider-took-5-trips-daily-4979437/

[2] IIITD. 2018. Delhi Open Transit Data. Retrieved Mar 7, 2019 from https://otd.delhi.gov.in/

[3] Outlook India. 2018. Delhi Is Planning Congestion Tax On Busy Roads, But Will It Work? Retrieved Mar 7, 2019 from https://www.outlookindia.com/website/story/congestion-tax-dear-mandarins-delhi-isnt-singapore-come-up-with-doable-idea/310363

[4] Times of India. 2018. Centre in poll hurry to kick off 4 infrastructure projects in NCR. Retrieved Mar 7, 2019 from https://timesofindia.indiatimes.com/city/delhi/centre-in-poll-hurry-to-kick-off-4-infra-projects/articleshow/68294617.cms

[5] Time of India. 2018. Traffic congestion costs four major Indian cities Rs 1.5 lakh crore a year. Retrieved Mar 7, 2019 from https://timesofindia.indiatimes.com/india/traffic-congestion-costs-four-major-indian-cities-rs-1-5-lakh-crore-a-year/articleshow/63918040.cms

[6] Economic Times. 2017. Taxi aggregator Uber picks up pace narrowing gap with Ola. Retrieved Mar 7, 2019 from https://economictimes.indiatimes.com/small-biz/startups/taxi-aggregator-uber-picks-up-pace-narrowing-gap-with-ola/articleshow/60049678.cms

[7] Economic Times. 2018. Speed bump: Uber, Ola face sharp slowdown in growth of rides in 2018. Retrieved Mar 7, 2019 from https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/uber-ola-enter-slow-lane-in-2018/articleshow/65927443.cms

[8] Hindustan Times. 2018. Car owners may have to pay to use 21 congested Delhi roads. Retrieved Mar 7, 2019 from https://www.hindustantimes.com/delhi-news/car-owners-may-have-to-pay-to-use-21-congested-delhi-roads/story-J65FdBNPg45FNN61CPgacL.html

[9] Uber. [n. d.]. How do ETAs work? Retrieved Mar 7, 2019 from https://help.uber.com/riders/article/how-do-etas-work?nodeId=6a4dbe13-0a86-4d11-a0ab-cc88f8171fb8

[10] Uber. 2017. Cancellation Fees in Delhi. Retrieved Mar 7, 2019 from https://www.uber.com/en-IN/blog/delhi/cancellation-fee/

[11] Uber. 2017. Cancellation Fees in India. Retrieved Mar 7, 2019 from https://www.uber.com/en-IN/blog/understanding-cancellation-fee/

[12] Uber. 2017. Global Cancellation Fees. Retrieved Mar 7, 2019 from https://help.uber.com/riders/article/am-i-charged-for-cancelling?nodeId=24e75a3b-cf44-44e4-abae-8c2dce3b07a3

[13] Dong Wang, Junbo Zhang, Wei Cao, Jian Li, and Yu Zheng. 2018. When Will You Arrive? Estimating Travel Time Based on Deep Neural Networks. In *AAAI*.

[14] Hanyuan Zhang, Hao Wu, Weiwei Sun, and Baihua Zheng. 2018. DEEPTRAVEL: a Neural Network Based Travel Time Estimation Model with Auxiliary Supervision. In *IJCAI*.

Data4UrbanMobility: Towards Holistic Data Analytics for Mobility Applications in Urban Regions

Nicolas Tempelmeier
L3S Research Center, Leibniz Universität Hannover
tempelmeier@L3S.de

Yannick Rietz
PROJEKTIONISTEN GmbH
rietz@projektionisten.de

Iryna Lishchuk
Institute for Legal Informatics, Leibniz Universität Hannover
iryna.lishchuk@iri.uni-hannover.de

Tina Kruegel
Institute for Legal Informatics, Leibniz Universität Hannover
kruegel@iri.uni-hannover.de

Olaf Mumm
Institute for Sustainable Urbanism, TU Braunschweig
o.mumm@tu-braunschweig.de

Vanessa Miriam Carlow
Institute for Sustainable Urbanism, TU Braunschweig
v.carlow@tu-braunschweig.de

Stefan Dietze
GESIS - Leibniz Institute for the Social Sciences
stefan.dietze@gesis.org

Elena Demidova
L3S Research Center, Leibniz Universität Hannover
demidova@L3S.de

ABSTRACT

With the increasing availability of mobility-related data, such as GPS-traces, Web queries and climate conditions, there is a growing demand to utilize this data to better understand and support urban mobility needs. However, data available from the individual actors, such as providers of information, navigation and transportation systems, is mostly restricted to isolated mobility modes, whereas holistic data analytics over integrated data sources is not sufficiently supported. In this paper we present our ongoing research in the context of holistic data analytics to support urban mobility applications in the Data4UrbanMobility (D4UM) project. First, we discuss challenges in urban mobility analytics and present the D4UM platform we are currently developing to facilitate holistic urban data analytics over integrated heterogeneous data sources along with the available data sources. Second, we present the MiC app - a tool we developed to complement available datasets with intermodal mobility data (i.e. data about journeys that involve more than one mode of mobility) using a citizen science approach. Finally, we present selected use cases and discuss our future work.

ACM Reference Format:
Nicolas Tempelmeier, Yannick Rietz, Iryna Lishchuk, Tina Kruegel, Olaf Mumm, Vanessa Miriam Carlow, Stefan Dietze, and Elena Demidova. 2019. Data4UrbanMobility: Towards Holistic Data Analytics for Mobility Applications in Urban Regions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3317055

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317055

1 INTRODUCTION

Contemporary urban mobility behavior is undergoing a rapid transition and paradigm shift and also is affected by a wide range of short-, medium- as well as long-term factors. These vary from aspects such as immediate climate conditions, regional construction sites or ephemeral events to long-term trends such as increasing negative environmental impacts and the widespread adoption of intermodal mobility chains (i.e. journeys involving several means of transportation and mobility such as walking, cycling, public transportation, etc.). Moreover, the lifestyle- and population-induced new demand for mobility leads to region-specific, ecological and traffic related problems in growing metropolitan areas and is a limiting factor for urban development.

Traditional urban traffic planning relies on complex models, combining trip generation, trip distribution, mode choice and route choice [21] to simulate travel demands. Data foundations of these models essentially consist of traffic census, socio-demographic data as well as infrastructural and service data, where large-scale and accurate data about actual mobility behavior, in particular in the context of intermodality, is costly to obtain and the aforementioned contextual factors are largely ignored.

Throughout the last decade, the widespread use of Web and mobile applications has led to an increasing availability of data, which captures both actual mobility needs and usage as well as contextual information about, for instance, traffic incidents, city events and weather conditions. This data has the potential to complement existing data sources and information systems currently used for handling urban mobility processes. In particular, within densely populated areas, the correlation of mobility behavior with data obtainable from mobility apps, public transportation websites and social media streams can uncover more complex dependencies and aid the development of supervised models which consider a wide variety of features and enable predictions of future needs.

Despite an overall increasing availability of datasets related to mobility behaviour, this data typically focuses on one mode of transportation, most prominently covering individual traffic and, to

some extent, public transportation services. Other modes of mobility that become particularly important in modern cities, e.g. cycling and walking, are typically not captured at scale. Furthermore, data sources coming from particular mobility services and transportation providers do not adequately capture intermodal mobility sequences. However, such data is of utmost importance to better understand and predict the actual demand in mobility and associated services.

The challenges related to collection and analysis of such data are manifold. They include the provision of tools and methods to capture and analyse intermodal mobility, designing incentives for city inhabitants to share their mobility data, protection of personal data to be collected in accordance to the legal framework as well as data analytics methods and models to provide added value for the individual participants, mobility service providers and city authorities.

In this work, we present our ongoing research within the *Data4UrbanMobility* project (D4UM)[1]. This research aims at gathering, augmenting and analysing mobility data in urban regions to address the problems presented above. We introduce the D4UM platform built to facilitate collection and analysis of mobility-related data from a variety of sources on a long term. In the project we in particular focus on the datasets available in the urban regions of Hanover, Wolfsburg and Brunswick (Germany), whereas the results will be transferable to other urban regions. The overall contributions of the project include a large annotated data catalogue including regionally and globally relevant mobility-related data sources, models built upon these data sources and analytic results in the particular regions.

Furthermore, the paper introduces the MiC-App - a human-centric data tool developed in the project framework to capture individual movements – tracks and modes – to complement existing sources in the context of intermodal urban mobility. Combined with other available data sources, the data collected through the MiC-App and the underlying D4UM platform facilitates the capturing and the analyses of data to better understand the growing demand in intermodal mobility in urban regions.

2 PROBLEM DESCRIPTION

The aim of the D4UM project is to facilitate efficient analytics and data-driven estimates of mobility demand and evaluation of mobility network quality, addressing the needs of different stakeholders. The stakeholders of the project include individual citizens, city administrations, mobility service providers and urban traffic planners. The aims of these stakeholders can be categorised with respect to the different time horizons, as follows:

- Short Term: Facilitate efficient access to mobility services, taking into account temporary high load (citizens).
- Medium Term: Provide new and optimise existing mobility services (mobility service providers).
- Long Term: Facilitate data-driven integrated urban planning processes (urban traffic planers, city administrations).

To address these problems the project develops the D4UM platform that facilitates collection and analytics of relevant data sources. The research questions that can be addressed using this platform include e.g.:

- Which external conditions such as climate or spatial structures influence the typical movement patterns of city inhabitants?
- How are the urban movement patterns influenced by the external factors, such as e.g. weather conditions, given specific urban contexts and mobility options?
- How can an increased demand in mobility services be determined and addressed?

3 D4UM PLATFORM

This section describes the architecture of the D4UM platform. This platform enables integrated analytics of heterogeneous data sources for different stakeholders in the context of urban mobility. Figure 1 provides an overview of the platform layers and its individual components.

The input layer of the D4UM platform consists of *heterogeneous data sources* that cover various aspects of urban mobility. These data sources are described in Section 4 in more detail.

The *data aggregation and integration* layer conducts all necessary pre-processing and transforms these data sources to comply with the D4UM data model. This data model formally specifies an integrated schema and establishes spatial, temporal and contextual connections across these sources, thus facilitating integrated analytics going across the dataset boundaries. For example, traffic speed records are aligned with the street segments obtained from the OpenStreetMap data.

The D4UM platform conducts long-term collection of mobility-related data from data streams to create an overview of relevant mobility data over longer time periods (i.e. months or years) within the *long-term data collection* component. In particular, recording of dedicated streaming sources such as traffic warnings published as RSS feeds or data extracted from social media channels (e.g. police channels on Twitter) allows for long-term analytics to observe patterns and temporal fluctuations, which would otherwise not be possible. *Data enrichment* facilitates collection of supplemental data by employing Web mining (e.g. by identifying events in social media streams) and citizen science tools (e.g. by collecting data with the MiC App presented in Section 5).

The *data analytics* layer introduces models that build on the integrated data. For example, analysis of long-term data collections can be employed to identify typical urban movement patterns. For instance, spatio-temporal dependencies between traffic conditions on different roads can reveal structural problems within the road network [6, 19]. Another example is the analysis of the impact of external factors (e.g. heavy rain or snowfall) on mobility behaviour [29]. Data-based forecasts can make use of long-term data collections to predict urban mobility patterns in the future. For example, in the presence of planned special events (e.g. football matches or concerts), an increased load on the mobility infrastructure such as roads and public transportation capacity can be estimated [27, 39].

Dedicated *interfaces* such as APIs and graphical user interfaces grant access to the data analytics results. APIs serve as a conceptual abstraction from the complex models and provide information in

[1] http://data4urbanmobility.l3s.uni-hannover.de/

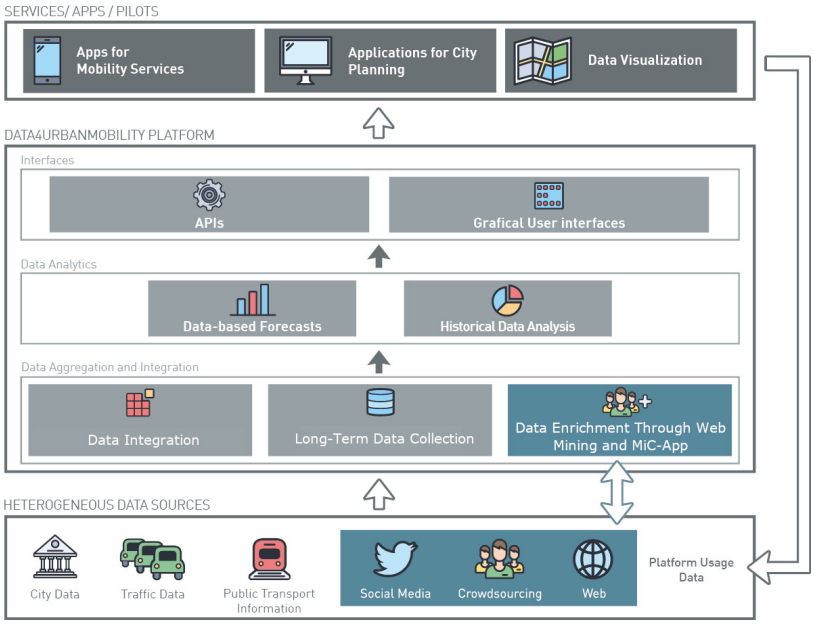


Figure 1: Overview of the Data4UrbanMobility platform architecture

a machine readable form, such as the GeoJSON[2] format. Graphical interfaces such as map layers graphically present the analysis results and allow for embedding a map into Web pages.

Finally the *services, apps and pilots* layer makes the analysis results available for end users. In particular, we develop a dashboard Web application that can be used by city planers to analyse traffic patterns on the long term, while citizens can use the MiC app to derive insights about their own mobility behaviour and voluntarily contribute data to the D4UM system.

4 INTEGRATION OF WEB-BASED MOBILITY DATA

An overview of existing data sources from which we currently extract regional and other relevant information in the context of urban mobility is presented in Table 1. The majority of these data sources (except the traffic flow information) are Web-based. Traffic Data information sources include traffic flow information and traffic feeds. Traffic flow data, provided as aggregated Floating Car Data (FCD), reflects the average speed of the road traffic with respect to the individual road segments, whereas traffic feeds contain traffic warnings and accident notifications provided as RSS feeds. Public Transport Information includes public transportation query logs and GTFS data. We obtain query log data from the *EFA*-system[3]. *EFA* is the official Web service for routing and time table information of public transportation for the region of Hannover. In this context, a query is a request for a public transportation route with specified origin, destination and departure time, issued via a Web-interface or a mobile application. Moreover, we consider Global Transit Feed Specification (GTFS)[4] data which provides timetable information for public transportation such as departure times, routes, etc. This data is complemented with regional weather conditions, which can potentially impact the selection of the transportation mode and routes.

Next to such directly related mobility data, we consider additional information obtained from social media and the Web. We consider social media data obtained from the Twitter streaming API[5], which potentially contains information about regional events as well as traffic incidents. Web data includes event-centric Web markup, which is prevalent in Web pages through standards such as RDFa[6], and Microdata[7]. We currently investigate in particular data complying with schema.org as the most established markup vocabulary on the Web so far [23]. In our previous work we developed methods to infer missing categorical information in noisy and sparse Web markup data [31], increasing usefulness of this data for event-centric applications. Furthermore, we consider event-centric focused crawls from the Web [7] and Web archives [8], [9], as well as Twitter data regarding events and traffic. Another source of event-centric information is the recently proposed EventKG knowledge graph [10, 11]. Finally, this information is complemented with geographic data (e.g. street networks, locations of event venues) obtained from *OpenStreetMap*[8]. The information contained in these sources is highly complementary.

These data sources can be used to address parts of the afore introduced research questions. For instance, in the context of events, the traffic flow and public transportation query data can be used to determine typical patterns while Web, Web archives, knowledge graphs and social media sources can provide information about

[2]https://tools.ietf.org/html/rfc7946
[3]https://www.efa.de
[4]https://developers.google.com/transit/gtfs/
[5]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[6]RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/
[7]http://www.w3.org/TR/microdata
[8]http://www.openstreetmap.org/

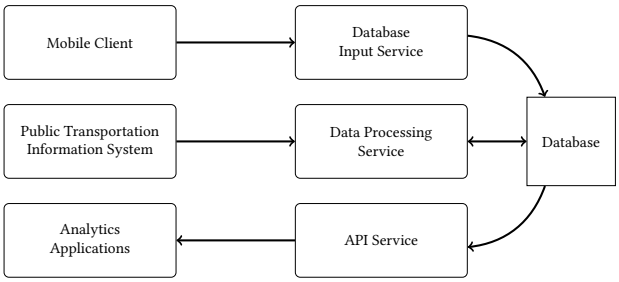


Figure 2: Overview of the MiC app data flow.

planned special events such as concerts or football matches. Another example is the computation of the average load on the roads with respect to the external factors, e.g. weather. These approaches require integration of data originating at several data sources, where semantic data descriptions, methods of dataset profiling [5], [4] and data quality analytics play an important role.

However, for comprehensive analysis of urban mobility some data is still missing. In particular foot walks, bicycle rides and intermodal data is not captured by any of the existing data sources. On the contrary, these modes become increasingly popular and thus important for urban mobility applications. To this end, the project introduces the MiC App, which can be used to collect the required data.

5 MIC APP

To address the lack of data regarding foot walks, bicycle rides and intermodal trips, the project introduces the *Move in the City* (MiC) app. This app poses a citizen science inspired method to gather movement data. Users of the MiC app can capture GPS traces of their trips with their mobile devices and voluntarily contribute the data to the D4UM project. In return, individual movement statistics are provided to the user. Moreover, the increased amount of available mobility data enables novel analytics of mobility behaviour. Based on the analytics, previously unseen problems can be identified and addressed, creating further benefits for the user and for the mobility ecosystem as a whole.

Movement data captured by the MiC app undergoes a preprocessing routine tailored to the specific needs of urban mobility analytics. Different modes of mobility such as walking, cycling, driving and the use of public transportation are automatically distinguished. Furthermore, if public transportation is used, the data is enriched with additional information such as entry stop, exit stop and the public transportation line used. This way, the data captured by the MiC app represents a valuable data source for further mobility analytics approaches.

5.1 Architecture

The architecture of the MiC app aims at keeping a low computation effort on the user device to reduce the energy consumption of the app. Furthermore, the architecture is platform independent to enable as many users as possible to use the app. To this end, MiC makes use of established, flexible Web technologies such as Web-based user interfaces and the MQTT protocol[12]. To further ensure platform independence, the app does not directly access sensor data of the mobile devices. Instead, the relatively new, built-in activity recognition APIs of the mobile operating systems are used. In addition to the activity recognition data, the app records fine grained location data.

Figure 2 provides an overview on the data flow. Data recorded by the mobile client is sent to a Web-based endpoint which stores the recorded data in a database. The data is then enriched with information about public transportation, obtained by querying the local information system for public transportation. Finally, a Web-based endpoint provides an API that makes the data accessible for data analytics applications.

5.2 App Design

MiC aims at motivating the users to use the app by providing an appealing and easy to use interface. Figure 3 presents two exemplary screenshots of the MiC app. The start view can be seen in Figure 3a. We achieve a lightweight user interaction by only requiring users to press one button to start or stop the recording of their movement. Figure 3b depicts the view presenting statistics of the individual user. The view presents the fraction of the user transportation modes by providing absolute and relative numbers as well as a graphical visualisation. Ideas for further user statistics include visualisation of the environmental impact of the users mobility behavior or the overall contribution to the urban mobility network. Figure 3c presents the map view, where the user can visualise captured trip data, i.e. GPS traces, on a map.

5.3 Piloting

The first closed test of the system was conducted by five participants who made 41 trips over the duration of five days, where the median duration of the recorded trips took 24 minutes. The participants wrote detailed travel diaries for these trips that were then used as a ground truth to assess the accuracy of the mode of travel classification. Due to limited GPS signals paths in underground trains, they were excluded from the study.

Table 2 presents the median duration and accuracy of transportation mode recognition of the recorded trips with respect to mode of transportation, where the total row summarises all recorded trips. Trips that include multiple modes of transportation were divided into individual trips with only one mode of transportation. In addition to classifying the correct transportation mode, we only considered tram and bus trips to be recognised correctly, if the correct entry stop, exit stop and public transportation line was identified by the system.

In total, 86.2% percent of the trips were recognized correctly by the system. The highest accuracy was achieved for the recognition of bicycle trips (100%). This is due to the relatively clear movement signature of riding a bicycle, which is well recognised by the activity recognition of the smartphones. The second highest accuracy is achieved for cars. Even though trams and busses achieve the lowest accuracy, the absolute accuracy is at least 73%. The reduced accuracy for the two classes is likely to be caused by the additional

[12] http://mqtt.org/

Table 1: An overview of regional and general mobility-relevant data sources for Lower Saxony, Germany.

Source	Description	Granularity / Size	Timespan	License	Format
Traffic Data					
Traffic flow	Average car traffic speed records in Lower Saxony, Germany.	Average traffic speed per road segment and time interval (15 min)	September 2017 - January 2019	Commercial	CSV
Traffic feeds	Traffic warnings and incidents[9].	~ 25000 notifications	June 2017-January 2019	Provider-specific	RSS feeds / XML
Public Transport Information					
Public transportation query logs	Query logs for public transportation routes and timetables.	$\sim 4 \cdot 10^6$ queries per month	October 2016-January 2019	Commercial	CSV
Global Transit Feed Specification (GTFS) data	Public transportation timetable information for Lower Saxony, Germany.	8800 stops, 2600 routes, $2 \cdot 10^6$ stop times	until January 2019	Open Data	CSV
City Data					
Rainfall data	Volume of rain in a region.	1 Record / hour / km^2	2005-2019	Open data	CSV
Social Media					
Twitter	Event- and location-centric tweets from Twitter API. Traffic information channels[10].	German tweets from Twitter API	May 2017-January 2019	Twitter API license	JSON
Web					
Event-centric Web markup	Annotated Web pages, e.g. using schema.org.	Web Data Commons event subset: 263×10^6 facts	until November 2017	Common Crawl ToU	RDFa, Micro-Data
Focused crawls	Event-centric crawls, news[11].	~ 22000 events located in Hannover, Germany	Crawl-specific	Provider-specific	HTML
EventKG	Multilingual event-centric temporal knowledge graph	$690 \cdot 10^3$ events and $2.3 \cdot 10^6$ temporal relations (V1.1)	until today	Creative Commons Attribution Share Alike 4.0	RDF
OpenStreetMap	Geometries (points, lines, polygons) annotated with properties. Subset for parts of Lower Saxony.	$2.6 \cdot 10^6$ facts	until January 2019	ODbL	XML

Table 2: Median duration and accuracy of mode of travel recognition with respect to transportation mode. Trips including multiple modes of transportation are divided into individual trips with only one mode of transportation.

Mode	Number	Median Duration	Accuracy
Bicycle	16	10 min.	100%
Car	14	9.5 min.	92.9%
Tram	13	7 min.	76.9%
Bus	15	12 min.	73.3%
Total	58	11 min.	86.2%

constraints, i.e. the recognition of start stop, entry stop and public transportation line.

The first public test phase of the system is currently ongoing. The application beta testing platforms by Apple and Google are being used to carry out the application. Users were recruited at universities and expositions. Table 3 presents statistics about data currently captured by the MiC app. To today, 78 participants were found to test the system under real-world conditions.

Table 3: Statistics of data currently captured by the MiC app.

Property	Value
Number of Users	78
Number of Recorded Trips	218
Average Trip Duration	69 min.
Number of Captured GPS Points	92.550

5.4 Data Protection

To ensure data protection compliant to latest regulations (e.g. the *General Data Protection Regulation* by the European Union), a concept was developed in joint work with researchers from legal informatics that includes the following components:

- The consent of app users (Article 6 GDPR);
- A privacy policy (Article 13 GDPR);
- Data pseudonymisation (Article 89 GDPR);
- The agreement between joint controllers (Article 26 GDPR).

The MiC app has an integrated privacy policy and consent form providing app users with the information about the data procession, data rights and possibility to consent. It contains details about the types of data the MiC app collects, the collection procedure, purposes of research, data storage and data sharing within the D4UM

(a) Start View **(b) User Statistics View** **(c) Map View**

Figure 3: Screenshots taken from the MiC app. Subfigure (a) presents the start view. Users can start recording their movement by pressing a single button. Subfigure (b) presents the user statistics view. It visualises details about the user specific mobility behaviour, e.g. the user specific fraction of transportation modes. Subfigure (c) presents the map view that projects captured trip data onto a map.

project, data rights of the users (i.a. access, deletion, withdrawal of consent) and contact details of the app providers.

The geo-referenced data collected by the MIC app is processed in pseudonymised form. By this, the data is purified from all personal identifiers, such as name, e-mail, IMEI number (a cellphone serial number), etc. Pseudonymisation is chosen as a de-identification measure (in contrast to anonymisation or irreversible de-identification) to enable enforcement of the data subjects' rights to access and deletion of the data, as may be requested by the users. This is part of an integrated and innovative citizen science approach which enables everyone to actively participate in research and planning processes with a strong effort to establish new standards for data sovereignty of citizens and cities.

6 USE CASES

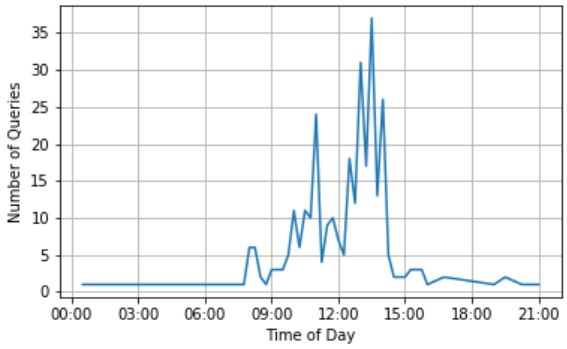
In this section we discuss two exemplary use cases for the D4UM platform and the MiC app.

6.1 Urban Mobility in Presence of Planned Special Events

Large-scale planned special events such as football games, concerts, etc. are known to have impact on both road traffic [15] and public transportation services [27]. Analysing mobility behaviour in such situations is particularly challenging due to a great number of influence factors (e.g. target audience, weather conditions, public transportation infrastructure). Moreover, large-scale events typically have an impact on all modes of transportation, which are mutually dependent. Thus, a holistic approach, which considers data about all modes of transportation as well as external factors, is required to gain a better understanding of the mobility behaviour in presence of planned special events.

We illustrate the complexity of the problem at the example of a football game that took place in the city of Hannover (Germany) on December 17th, 2017 at 3:30pm. Figure 4 presents the conditions around the football stadium in Hannover for both road traffic and public transportation services. Figure 4a depicts the road traffic conditions around the stadium half an hour before the start of the game. As we can observe, the roads with a high load (red color) as well as the roads with good traffic conditions (green color) are present nearby the stadium. This illustrates that the impact of planned special events might be complex and such impact does not necessarily evenly spread around the event venue, but might be subject to other factors such as road topology or availability of parking spaces. Figure 4b presents the number of queries to the public transportation information system that were issued for the bus stop near the stadium during the day of the football game. We observe a relatively low number of queries during the morning hours. At 9am, the number of queries starts to rise and reaches their maximum at 2:30pm, one hour before the game starts. We assume that an increased number of queries in the temporal proximity of the event is likely related to the football game. Overall, we observe,

(a) Traffic conditions on the major roads located nearby the football stadium half an hour before the game starts. Red color indicates heavy load, green color indicates low load.



(b) Number of public transportation queries for the bus stop nearest to the football stadium with respect to the time of day on the day of the football game.

Figure 4: Illustration of an impact of a football game on road traffic and public transportation. The football game took place on December 17th, 2017 at 3:30pm in Hannover, Germany.

that a single factor, i.e. the taking place of a large-scale event, can impact both road traffic and public transportation simultaneously. Therefore, analysis of these effects should take both mobility modes into account. To this extent, the D4UM platform presented in this paper aims at holistic data analysis of urban mobility data.

6.2 Analytics of Urban Mobility Infrastructure

The rapid growth of cities has lead to an increased demand of suitable urban mobility infrastructure. New services such as car sharing, (e-)bike sharing and ride sharing have emerged and begin to gain importance within the mobility services ecosystem. However, data about these modes of transportation is typically only sporadically available. Moreover, intermodal data which captures different modes of transportation on a single journey is evenly rare.

The interdependencies between individual and public transportation increase the complexity of the problem. For instance, a poor coverage of public transportation services might lead to an increased load on the road infrastructure. In turn, an overloaded road system can lead to an increased demand of public transportation services or bicycle tracks.

Therefore it is crucial for city planers and mobility service providers to conduct holistic analysis, which take all modes of transportation into account. Such analysis should include the identification of:

- typical mobility patterns
- demand of mobility services
- interdependencies between transportation modes
- coverage of public transportation services

The D4UM platform enables the holistic analysis of urban mobility data in these scenarios. Furthermore, the MiC app presented in this paper is a valuable approach to capture rarely available data about urban mobility behaviour, including bicycle rides and intermodal data and complement available data sources.

7 RELATED WORK

In this section we discuss related work in the area of smart city mobility systems as well as approaches to predict individual aspects of urban mobility.

7.1 Smart City Mobility Systems

[24] conducted a systematic review of smart city data analysis approaches and provide taxonomies for data sources, data analytic methods and smart city services. Urban CPS [38] is a cyber physical system that integrates floating car data, cellphone data and public transportation data to make prediction about real time traffic speeds. [17] employ semantic technologies to develop STAR-CITY, a system for traffic prediction and reasoning, used for spatio-temporal analysis of the traffic status as well as for the exploration of contextual information such as nearby events. [3] presents a system that makes use of cellphone data to analyse the demand of public transportation services within a city. While these approaches focus on predictions for one mode of mobility only and mainly use a single dataset as their primarily data source, we consider the other modes of transportation as well, e.g. bicycle rides or public transportation. [1] provides a summary of common enterprise, logical and physical architectures for digital smart city applications. While the authors consider architectures for general smart city applications, we focus on the mobility domain. [33] proposed the Compressed Start-End Tree (CSE-tree), a spatio-temporal index structure that can be used to effectively index and retrieve temporal GPS data which is in particular relevant for smart city mobility applications.

7.2 Urban Mobility Analytics

Recently a number of studies addressed several individual prediction tasks at the interface of the urban infrastructure and mobility. We consider these approaches as potential use cases for the D4UM platform.

Traffic Forecasting: Short-term forecasting of urban road traffic has been the focus of numerous studies where data sparsity is a particular challenge. [35] tackles the problem by using sparse FCD (floating car data) and a context-aware tensor decomposition approach to estimate travel times for road segments for which no FCD is available. The information is then used to estimate the required travel time for a given route. [22] proposed a framework for the city-wide inference of traffic volume. They make use of a semi-supervised learning algorithm that can be used with sparse loop detector data as well as taxi GPS data. Similar, [34] employs a hidden Markov model to estimate traffic speeds of a road network based on sparse FCD where the speed to be estimated on a single road is considered as a hidden state.

Social Network Data: Further approaches utilized location-based social network (LBSN) data. [28] leverages LBSN data to identify functional urban regions. They employ latent Dirichlet allocation and unsupervised machine learning algorithms to determine the regions. [18] investigated the general predictability of LBSN data. They conducted a case study on Foursquare datasets, where users indicate their geographic location, i.e. the users can indicate that they are at a certain event venue. The authors do not focus on a specialised prediction task, but provide general insights on working with the aforementioned data. [36] make use of LBSN data to infer boundaries of functional regions in urban environments. They construct a mobility network from spatial user interactions and delineate boundaries by identifying strongly connected communities within the network space. [25] detects events from social media by employing a hashtag-based algorithm. The event information extracted from the social media is then used for prediction of the public transportation flow.

Structural Analysis: Another class of approaches focuses on discovering structural dependencies within cities by analysing urban mobility data. [2] propose a method to keep track of the congestions in urban road networks to identify unstable road segments. [13] make use of context-aware tensor decomposition to identify so called *urban black holes*, i.e. traffic anomalies with a greater inflow than outflow. [12] detect urban black holes by using a grid-based index-structure that is build on top of a spatio-temporal graph representing the road network. They extract candidate cells from the index which then are used to determine the exact subgraphs that are urban black holes. [19] identify cascading patterns on congested roads. They propose a generative probabilistic model that maximises the likelihood of a cascade to be present with respect to the observed traffic data. [14] employs topic modelling to analyse urban street works. They proposed the concept of interactional regions, i.e. regions that commonly bound routes within the street networks. [32] leverage taxi flow data to learn vector representation of city regions. The representations are then used to make predictions about the regions such, e.g. crime rate, average income or average house prices. [20] employ deep learning techniques, i.e. a combination of restricted Boltzmann machines and recurrent neural networks, to learn high-dimensional congestion patterns from taxi GPS data. [26] make use of taxi GPS-trajectories to classify the land use of urban areas. They propose an iterative DBSCAN algorithm to cluster regions with respect to the frequency with which passengers are picked up or set down. They make use of the same information to classify the land use of regions, e.g. the land use for hospitals or commercial districts. [37] proposes the infinite urbanization process model that employs a topic modelling approach to simultaneously discover the function of an urban region (i.e. the distribution of present shops, restaurants, etc.) and to estimate the region popularity (e.g. in terms of real estate prices). Similar, [30] extracts urban regions of interest from online map search queries. They propose a spatio-temporal latent factor model which identifies travel patterns that influence points of interest.

Special Traffic Conditions Several approaches target the identification of problematic segments and areas of urban networks under specific conditions (e.g. planed special events). [16] propose the use of an artificial neuronal network to identify road segments that are typically affected by planned special events that take place at a particular venue. [39] proposed an approach to detect events from traffic data. They make use of a two-dimensional grid to partition the space. The authors then infer a graph-based representation of the grid that captures the flow of vehicles between the individual cells of the grid where the root of the graph is located at the events location. Finally, [27] investigate the effect of public events on the public transportation network. They propose a Bayesian additive model that can be employed to gain an understanding of public transportation demand in the presence of events. I.e. the model is able to predict the number of public transportation trips to the venues where the respective event takes place.

8 CONCLUSIONS & FUTURE WORK

In this paper we presented our ongoing work towards holistic urban data analytics conducted in the context of the Data4UrbanMobility project. We presented the D4UM platform that facilitates seamless long-term analytics of heterogeneous mobility-related data sources, including but not limited to floating car data, weather conditions, traffic warnings and Web queries. Furthermore, we presented the MiC app - a citizen science application that facilitates complementing this data with intermodal mobility patterns of city inhabitants. In our future work we intend to further increase the MiC app user base, and implement further use cases on top of the integrated D4UM platform.

ACKNOWLEDGMENTS

This work was partially funded by the Federal Ministry of Education and Research (BMBF), Germany, "Data4UrbanMobility" project, grant ID 02K15A040.

REFERENCES

[1] L. Anthopoulos and P. Fitsilis. 2010. From Digital to Ubiquitous Cities: Defining a Common Architecture for Urban Development. In *Proceedings of the 2010 Sixth International Conference on Intelligent Environments*. 301–306.
[2] Tarique Anwar, Chengfei Liu, Hai L. Vu, and Md. Saiful Islam. 2016. Tracking the Evolution of Congestion in Dynamic Urban Road Networks. In *Proceedings of the ACM CIKM 2016*.
[3] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. 2013. AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). 663–666.
[4] Stefan Dietze, Elena Demidova, and Konstantin Todorov. 2019. RDF Dataset Profiling. In *Encyclopedia of Big Data Technologies*.
[5] Mohamed Ben Ellefi, Zohra Bellahsene, John G. Breslin, Elena Demidova, Stefan Dietze, Julian Szymanski, and Konstantin Todorov. 2018. RDF dataset profiling - a survey of features, methods, vocabularies and applications. *Semantic Web* 9, 5 (2018), 677–705.

[6] U. Feuerhake, O. Wage, M. Sester, N. Tempelmeier, W. Nejdl, and E. Demidova. 2018. Identification of similarities and prediction of unknown features in an urban street network. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4 (2018), 185–192.

[7] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2015. iCrawl: Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling. In *Proceedings of the JCDL'15*.

[8] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2017. Extracting Event-Centric Document Collections from Large-Scale Web Archives. In *Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017*. 116–127.

[9] Gerhard Gossen, Thomas Risse, and Elena Demidova. 2018. Towards extracting event-centric collections from Web archives. *International Journal on Digital Libraries* (27 Oct 2018).

[10] Simon Gottschalk and Elena Demidova. 2018. EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In *Proceedings of the 15th International Conference, ESWC 2018*. 272–287.

[11] Simon Gottschalk and Elena Demidova. 2019. EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation. *Semantic Web* (2019).

[12] Liang Hong, Yu Zheng, Duncan Yung, Jingbo Shang, and Lei Zou. 2015. Detecting Urban Black Holes Based on Human Mobility Data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, Article 35, 10 pages.

[13] Li Jin, Zhuonan Feng, and Ling Feng. 2016. A Context-aware Collaborative Filtering Approach for Urban Black Holes Detection. In *Proceedings of the ACM CIKM 2016*.

[14] Kira Kempinska, Paul Longley, and John Shawe-Taylor. 2018. Interactional regions in cities: making sense of flows across networked systems. *International Journal of Geographical Information Science* 32, 7 (2018), 1348–1367.

[15] Simon Kwoczek, Sergio Di Martino, and Wolfgang Nejdl. 2014. Predicting and visualizing traffic congestion in the presence of planned special events. *Journal of Visual Languages & Computing* 25, 6 (2014), 973 – 980.

[16] Simon Kwoczek, Sergio Di Martino, and Wolfgang Nejdl. 2015. Stuck Around the Stadium? An Approach to Identify Road Segments Affected by Planned Special Events. In *IEEE ITSC 2015*.

[17] Freddy Lécué, Simone Tallevi-Diotallevi, Jer Hayes, Robert Tucker, Veli Bicer, Marco Luca Sbodio, and Pierpaolo Tommasi. 2014. STAR-CITY: Semantic Traffic Analytics and Reasoning for CITY. In *Proceedings of the IUI '14*.

[18] Ming Li, René Westerholt, Hongchao Fan, and Alexander Zipf. 2018. Assessing spatiotemporal predictability of LBSN: a case study of three Foursquare datasets. *GeoInformatica* 22, 3 (2018), 541–561.

[19] Yuxuan Liang, Zhongyuan Jiang, and Yu Zheng. 2017. Inferring Traffic Cascading Patterns. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2017, Redondo Beach, CA, USA, November 7-10, 2017*. 2:1–2:10.

[20] Xiaolei Ma, Haiyang Yu, Yunpeng Wang, and Yinhai Wang. 2015. Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. *PLOS ONE* 10, 3 (03 2015), 1–17.

[21] Michael McNally. 2008. The four step model. In *Handbook of Transport Modeling*, David A. Hensher and Kenneth J. Button (Eds.). Emerald, Bingley, 35–53.

[22] Chuishi Meng, Xiuwen Yi, Lu Su, Jing Gao, and Yu Zheng. 2017. City-wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2017*. 1:1–1:10.

[23] Robert Meusel, Petar Petrovski, and Christian Bizer. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In *Proceedings of the ISWC'14*.

[24] Vaia Moustaka, Athena Vakali, and Leonidas G. Anthopoulos. 2018. A Systematic Review for Smart City Data Analytics. *ACM Comput. Surv.* 51, 5, Article 103 (Dec. 2018), 41 pages. https://doi.org/10.1145/3239566

[25] Ming Ni, Qing He, and Jing Gao. 2017. Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Trans. Intelligent Transportation Systems* 18, 6 (2017), 1623–1632.

[26] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li. 2013. Land-Use Classification Using Taxi GPS Traces. *IEEE Transactions on Intelligent Transportation Systems* 14, 1 (2013), 113–123.

[27] F. Rodrigues, S. S. Borysov, B. Ribeiro, and F. C. Pereira. 2017. A Bayesian Additive Model for Understanding Public Transport Usage in Special Events. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 11 (2017).

[28] Gao Song, Janowicz Krzysztof, and Couclelis Helen. 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21, 3 (2017), 446–467.

[29] R. Soua, A. Koesdwiady, and F. Karray. 2016. Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory. In *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*. 3195–3202.

[30] Ying Sun, Hengshu Zhu, Fuzhen Zhuang, Jingjing Gu, and Qing He. 2018. Exploring the Urban Region-of-Interest Through the Analysis of Online Map Search Queries. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 2269–2278.

[31] Nicolas Tempelmeier, Elena Demidova, and Stefan Dietze. 2018. Inferring Missing Categorical Information in Noisy and Sparse Web Markup. In *Proceedings of The Web Confonference 2018*.

[32] Hongjian Wang and Zhenhui Li. 2017. Region Representation Learning via Mobility Flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, 237–246.

[33] L. Wang, Y. Zheng, X. Xie, and W. Ma. 2008. A Flexible Spatio-Temporal Indexing Scheme for Large-Scale GPS Track Retrieval. In *Proceedings of the Ninth International Conference on Mobile Data Management (mdm 2008)*. 1–8. https://doi.org/10.1109/MDM.2008.24

[34] Xiaomeng Wang, Ling Peng, Tianhe Chi, Mengzhu Li, Xiaojing Yao, and Jing Shao. 2016. A Hidden Markov Model for Urban-Scale Traffic Estimation Using Floating Car Data. *PLOS ONE* 10, 12 (12 2016), 1–20.

[35] Yilun Wang, Yu Zheng, and Yexiang Xue. 2014. Travel Time Estimation of a Path Using Sparse Trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 25–34.

[36] Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science* 31, 7 (2017), 1293–1313.

[37] Bang Zhang, Lelin Zhang, Ting Guo, Yang Wang, and Fang Chen. 2018. Simultaneous Urban Region Function Discovery and Popularity Estimation via an Infinite Urbanization Process Model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 2692–2700.

[38] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. 2015. UrbanCPS: A Cyber-physical System Based on Multi-source Big Infrastructure Data for Heterogeneous Model Integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems (ICCPS '15)*. 238–247.

[39] Xun Zhou, Amin Vahedian Khezerlou, Alex X. Liu, M. Zubair Shafiq, and Fan Zhang. 2016. A traffic flow approach to early detection of gathering events. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2016*. 4:1–4:10.

CyberSafety 2019 Chairs' Welcome Message

We are excited to offer CyberSafety 2019, the Fourth Workshop on Computational Methods in Online Misbehavior, affiliated with The Web Conference 2019. This workshop offers a forum for researchers in academia, industry and government to share novel research on topics broadly related to the urgent theme of cybersafety, devoted to making the Web safer for its users. This is an area of exploding interest as abuse vectors such as social bots, online misinformation and propaganda, user deception and scams, harassment, hate speech, cyberthreats, cyberbullying and more have proliferated in recent years. A particular aim of the workshop is to encourage interdisciplinary research that brings together computer scientists and researchers from other disciplines, and provides a forum for these researchers to share their ideas.

This year, we received 23 paper submissions, all of which were reviewed by the Technical Program Committee (TPC). Based on reviews, the TPC Co-Chairs selected 8 papers for full presentation, and 6 additional papers for short presentation at the workshop, spanning topics that include spam and bot detection, misinformation, harassment, cyberbullying and more. In addition, the workshop will include several keynote talks from industrial and academic speakers on modern challenges and solutions in computational misbehavior. A detailed program is available at https://cybersafety2019.github.io/ for more details.

This iteration of the workshop succeeds three prior ones, co-located with

- WWW 2018 - https://cybersafety2018.github.io/
- WWW 2017 - https://cybersafety2017.github.io/
- CIKM 2016 - https://cybersafety2016.github.io/

Sincerely,

Richard Han and Neil Shah
TPC Co-Chairs, CyberSafety 2019

Homa Hosseinmardi, Srijan Kumar, and Qin Lv
Workshop Co-Chairs, CyberSafety 2019

CyberSafety 2019 Workshop Organization

Organizing Committee

- Homa Hosseinmardi, University of Southern California, USA
- Srijan Kumar, Stanford University, USA
- Qin Lv, University of Colorado Boulder, USA

TPC Co-Chairs

- Richard Han, University of Colorado Boulder, USA
- Neil Shah, Snap Research, USA

Steering Committee

- Jeremy Blackburn, University of Alabama at Birmingham, USA
- Bert Huang, Virginia Tech, USA
- Shivakant Mishra, University of Colorado Boulder, USA

Publicity Co-Chairs

- Rahat Ibn Rafiq, University of Colorado Boulder, USA
- Michael Sirivianos, Cyprus University of Technology, Cyprus

Program Committee

- James Caverlee, Texas A&M University, USA
- Emiliano De Cristofaro, University College London, UK
- April Edwards, Elmhurst College, USA
- Emilio Ferrara, University of Southern California USA
- Peng Gao, Princeton University, USA
- Meng Jiang, University of Notre Dame, USA
- Haewoon Kwak, Qatar Computing Research Institute, USA
- Nicolas Kourtellis, Telefonica Research, Spain
- Evangelos Papalexakis, University of California Riverside, USA
- Elaheh Raisi, Virginia Tech, USA
- Zubair Shafiq, University of Iowa, USA
- Michael Sirivianos, Cyprus University of Technology, Cyprus
- Francesca Spezzano, Boise State University, USA
- Gianluca Stringhini, Boston University, USA
- Mingyi Zhao, Snap Inc, USA

Detect Me If You Can: Spam Bot Detection Using Inductive Representation Learning

Seyed Ali Alhosseini
Hasso-Plattner-Institute
University of Potsdam
Potsdam, Germany
seyedali.alhosseini@hpi.de

Raad Bin Tareaf
Hasso-Plattner-Institute
University of Potsdam
Potsdam, Germany
raad.bintareaf@hpi.de

Pejman Najafi
Hasso-Plattner-Institute
University of Potsdam
Potsdam, Germany
pejman.najafi@hpi.de

Christoph Meinel
Hasso-Plattner-Institute
University of Potsdam
Potsdam, Germany
christoph.meinel@hpi.de

ABSTRACT

Spam Bots have become a threat to online social networks with their malicious behavior, posting misinformation messages and influencing online platforms to fulfill their motives. As spam bots have become more advanced over time, creating algorithms to identify bots remains an open challenge. Learning low-dimensional embeddings for nodes in graph structured data has proven to be useful in various domains. In this paper, we propose a model based on graph convolutional neural networks (GCNN) for spam bot detection. Our hypothesis is that to better detect spam bots, in addition to defining a features set, the social graph must also be taken into consideration. GCNNs are able to leverage both the features of a node and aggregate the features of a node's neighborhood. We compare our approach, with two methods that work solely on a features set and on the structure of the graph. To our knowledge, this work is the first attempt of using graph convolutional neural networks in spam bot detection.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Security and privacy** → *Social network security and privacy*; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Social Media Analysis, Bot Detection, Graph Embedding, Graph Convolutional Neural Networks

ACM Reference Format:
Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. 2019. Detect Me If You Can: Spam Bot Detection Using Inductive Representation Learning. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316504

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316504

1 INTRODUCTION

Online Social Networks (OSN) have provided a means of communication for individuals to share information and express their opinions in a free and simple manner. Twitter, Facebook and other social media websites have changed the way we consume news and interact with one another. The important role of these platforms has resulted in attempts by interest groups to influence users, seize their attention and ultimately change public opinion [3, 11].

Social Bots are automated user accounts operated by a computer program mimicking human behaviour with the intention of abusing the social media platform [3, 6]. They have evidently become a threat to online social networks with their malicious behavior spamming with advertisement and scam URLs, promoting a specific hashtag, spreading misinformation and impacting elections.

The research community has proposed several approaches for bot detection. The difference in these work varies depending on the definition of a bot account, the selected feature set representing accounts and the machine-learning algorithm used for classifying bot accounts from normal user accounts.

However, spam bot detection remains an open challenge for several reasons. The first reason relies in the definition of bot accounts as there is no single definition to precisely determine an account as a bot account. This is an important matter specially for building a ground truth dataset. Another issue as reported by [3, 17] is that bots have become more advanced and sophisticated in avoiding the existing proposed detection methods. In fact, bots have been evolving over time. [6] has given attention to the rise of social bots that are designed to emulate human-like behavior. The social bots are able to interact with other accounts, post tweets in different topics, and display a similar activity like humans [3].

Recent advances in deep learning for graph-structured data has led to a new method of representation learning named Graph Convolution Networks (GCNs) [9, 10]. The main idea of GCNs is to represent a node in a vector space based on its features and the features of its neighboring nodes using neural networks. The advantage of GCNs is that it captures both the node features and the graph structure to learn a low-dimensional representation of nodes [7, 8].

In this work we propose an inductive representation learning approach for bot detection based on the user profile features and the social network graph. The main contributions of this work are summarized as follows:

- We deploy graph convolutional neural networks on a well-known spam bots dataset previously used in the literature.
- We compare our approach with two algorithms by a MLP classifier and applying the Belief Propagation on the dataset.
- We show that using the graph structure in our method gains better performance in spambot detection.

The remainder of this paper is structured as follows. First we cover the previous related work in spam bot detection and graph convolutional neural networks. Section 3 describes in detail the dataset used. In section 4, we provide an overview of our methodology. We illustrate are results in section 5 and discuss the limitations of our work and suggestions for future work. Finally, we conclude this paper in section 6.

2 RELATED WORK

In this section, we first review the literature on spambot detection and compare each work by their definition for spambots, the features used and the classification algorithm they employed. Next, we look at graph convolutional networks.

[12] proposed a method working as a honeypot trap for bot accounts. They created 60 twitter accounts and started posting meaningless tweets that would have no interest for humans. Despite this fact, they were able to draw some accounts' attention to follow the accounts they made. Analyzing these accounts in detail showed that they were in fact bot accounts trying to increase their following list.

[17, 18] used a conservative definition for bot accounts considering only accounts who post URLs linking to malicious content. They also introduced and considered several robust features on the BayesNet classifier to predict spam accounts. Yang and et al investigated the different approaches bots take for avoiding detection by Twitter. Their findings show that bots tend to increase the reputation of their accounts by purchasing followers and posting more tweets.

[4] introduced a DNA-inspired technique that models each account as a sequence of behavioral information and detects spambots based on similar sequences. They categorized each users' tweets into different types and based on whether a tweet contains URLs, hashtags, pictures, etc. it will be assigned a different character. The similarity of the accounts is measured by the longest common substring in their DNA sequences.

BotOrNot [5] used the random forest classifier algorithm on more than 1000 features to detect bots. The features are categorized in 6 groups: network (degree distribution, clustering coefficient, ...), users' account information, friends (number of followers, followings, ...), temporal (tweet rate, ...), content (natural language processing, ...) and sentiment features. The downside of BotOrNot is that it was trained on English tweets so its performance declines on bots which are tweeting in another language than English.

DeBot [1, 2] is an unsupervised bot detection system. The idea behind their work is that accounts with a high correlation in their activities (tweet, retweet, ...) have a high chance of being bots. DeBot monitors the activities of accounts over a specific period and creates a time series for each account. It then clusters accounts based on the similarity of their time series using a lag-sensitive hashing method. Finally, DeBot reports the accounts with a high correlation as bots.

[13] defined spam bots as content polluters that try to take over a discussion for political or advertising reasons. Their approach considers individual tweets for detecting bots. Instead of focusing on the friend and follower network, they created the event network where the nodes are the users and the edges are based on users having tweeted on the same event. They also compute the diversity of a tweet based on the URLs and hashtags it has mentioned. Results of their work indicate that spam bots operate as a group often tweeting at the same time.

2.1 Graph Convolutional Networks

Graph structures are used in many domains and applications such as social networks, recommender systems etc. The challenging task for graph structures is how to use them in machine learning algorithms. The initial works in this area considered the statistical data of the graph like the degree of the nodes, the centrality and betweenness coefficients as features for training models. In other words, they considered the graph structure as a pre-processing step to extract structural information. Therefore, these approaches do not use the graph structure in the learning phase. Another downgrade for these approaches is that computing the graph statistics has high complexity and the output of it cannot be used on unseen data.

With the recent advances in Convolutional Neural Network (CNN) there have been efforts to adapt this popular deep learning model for encoding graph structures. Two main approaches have been used for embedding the graph structure into a dimensional space. The difference between these approaches is based on how the convolution operation is defined. The first approaches aim to take a fixed-length node sequence of the graph structure and directly use it in the original CNN models that work in the Euclidean domain. Alternatively, the other methods model the graph structure to non-Euclidean domains. [10] proposed graph convolutional network (GCN); that considers spectral convolutions on graph structures. The term convolutional is used since a node's neighborhood is considered as its representation. Their method can be considered as the initial steps for graph semi-supervised classification tasks. However, the drawbacks of their approach are that it requires the full graph Laplacian to be calculated and the output embeddings of a node in each layer are dependent on all its neighbors at the previous layer.

Most recently [8] introduced GraphSage; a node embedding algorithm that uses neural networks to learn embeddings for nodes in the graph structure. Their main contribution is that they solve the limitations mentioned above and show how to aggregate information from a node's neighborhood. Their method consists of two main phases:

(1) **Defining the computation graph and training the neural nets**
The structure of a node's neighborhood will define the computation graph for training the neural networks. In this phase, the objective is to build neural networks that will

ensure nodes close to each other have similar embeddings while nodes far from one another have different embeddings.

(2) **Propagation** For each node, the information of its neighbors is aggregated and passed through the neural networks trained in the first phase.

3 DATASET

There are several well-known datasets collected by different research groups specifically for bot detection on Twitter. Lee et al. [12] provide a social honeypot dataset that contains approximately 22000 content polluters. They have gathered the accounts' metadata and tweets of each account. However, in their released dataset they have anonymized the Twitter account ids. Therefore collecting further information is not possible. Cresci et al. have worked on different Twitter datasets in [3] and by using a crowdsourcing platform they labeled the different types of accounts. [16] released the twitter ids of the accounts they detected as spambots.

Yang et al. 2013 collected Twitter spammers and their dataset contains each account's followers and followings. To the best of our knowledge, this is the dataset we found which has gathered this information for the Twitter accounts. The authors of that work have kindly shared their dataset and we have used the dataset in the present paper. The dataset contains 11000 nodes and 2342816 edges between them.

Table 1 shows the statistics of the dataset used in this paper. The age, tweets and neighbors columns indicate the average amount in each group. The age column is the average age of the accounts reported in days and normalized by setting the oldest day as the first day. The majority of edges between nodes are user to user connections. However, around 5.4% of the edge relations include bot accounts.

	Accounts	Age	Tweets	Neighbors
bots	1000	3023.80	220.90	1963.84
users	10000	3174.28	4658.52	21579.76
relation	bot-bot	bot-user	user-bot	user-user
	2673	73363	50153	2216627
	0.11%	3.13%	2.14%	94.61%

Table 1: Dataset statistics

Figure 1 shows the degree distribution of the accounts in the dataset. Most accounts have a small number of followers and followings and there are a few accounts which have more than 1000 accounts in their neighborhood.

Figure 2.a shows the age and the length of user account name for both bots and users accounts. As shown in figure 2.b and reported in previous work [6] bot accounts have smaller age meaning they were created more recently compared to user accounts. Also as [13] indicated there is no significant difference in length of the accounts name.

4 METHODOLOGY

We used an inductive representation learning approach similar to [8, 9] for detecting twitter bot accounts.

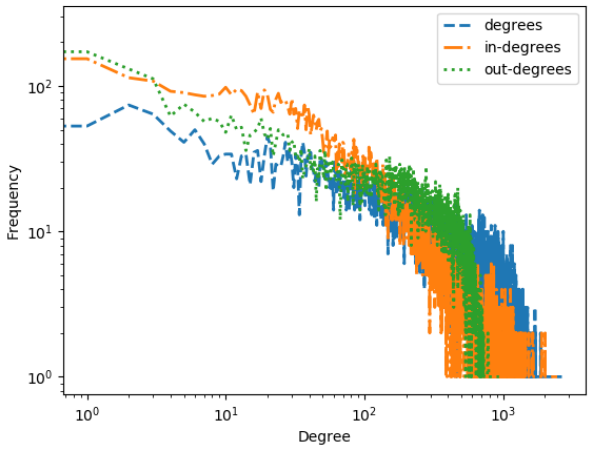


Figure 1: The degree distribution of the nodes in graph. The figure is drawn in log-log scale.

4.1 Problem definition

Let $G = (V, E)$ be a graph where for each $v \in V$ exists a feature vector X_v and a binary label $y \in \{0, 1\}$ associated with it. The goal is to find an embedding vector h_v for each node $v \in V$ such that $f(h_v)$ predicts the label of the node in the graph.

Similar to convolution filters in image processing, graph convolutional networks consider the attributes of a node's neighbors as a representation for that node. Let us define k as the depth of the neighbors of a node from which information is aggregated. If k=1 only the information from its own neighbors will be considered. For k=2 the information is gathered also from the neighbors of its neighbors and so on. The output h_v^k at each depth is calculated as follows:

$$h_{N(v)}^k = mean(\{h_u^{k-1}, \forall u \in N(v)\}) \quad (1)$$

$$h_v^k = f^k(h_v^{k-1}, h_{N(v)}^k) = \sigma(W^k \cdot concat(h_v^{k-1}, h_{N(v)}^k)) \quad (2)$$

Where $h_{N(v)}^k$ is the average of the embedding vectors from v's neighbors. h_v^k is the output which is concatenated with v's previous embedding.

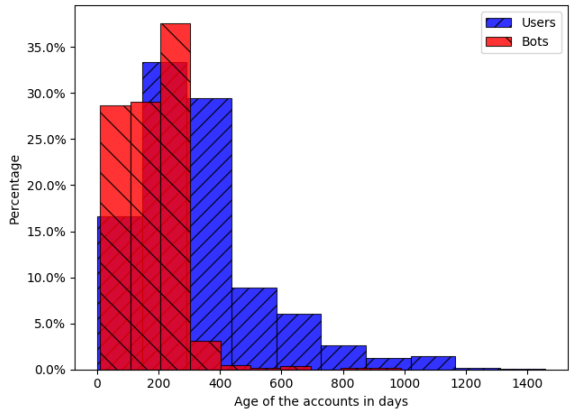
The neural networks are optimized based on the cross-entropy loss function:

$$J(f^k(h_v^{k-1}, h_{N(v)}^k), y) = \\ -\sum_{v \in V} y \log(f(X_v)) + (1 - y) \log(1 - f(X_v)) \quad (3)$$

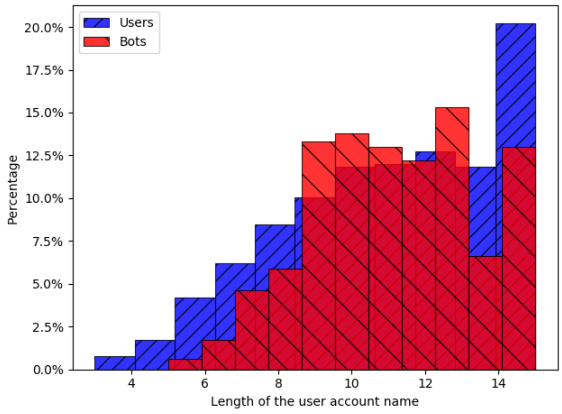
4.2 Features

The initial vector (X_v) for each user consists of the features that can be retrieved directly from the Twitter API[1]. The feature vector consists of :

[1]https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object.html.



Age (a)



Account Length Name (b)

Figure 2: Bots(red) and Users(blue) attributes

	Feature Name	Description
1	age	The created_at attribute returns the datetime that an account was created on Twitter. The age feature is computed by the number of days from the created_at date.
2	favourites_count	This feature indicates the number of tweets a user has liked.
3	statuses_count	The number of tweets including the retweets a user has posted.
4	account length name	The length of an account's name
5	followers_count	followers_count shows the number of follower an accounts has.
6	friends_count	The friends_count attribute shows the number of accounts the user is following.

Table 2: Features

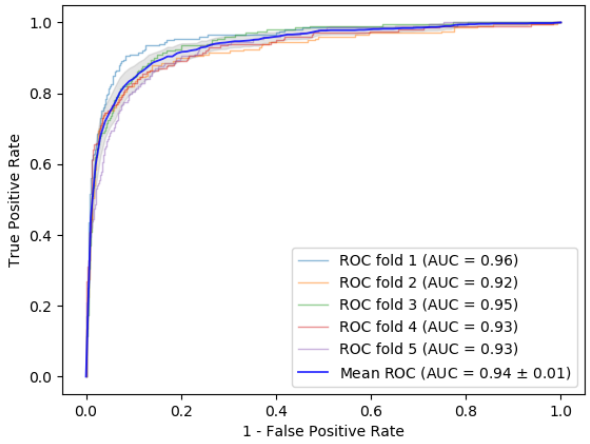


Figure 3: ROC curve over 5-fold cross-validation

5 EVALUATION

In this section, we evaluate the performance of our approach. We conducted a 5-fold cross-validation on the dataset to evaluate the accuracy of the model. Figure 3 shows the area under curve for each fold. On average the GCNN has 0.94 accuracy measured by the area under roc curve.

We measured the precision, recall and f1 metrics as shown in Table 5 for the evaluation. Choosing a meaningful evaluation metric for the classification task is important. For example, it is possible to use the precision measure defined in equation 4 to evaluate the performance of a model. In this case, the measures are calculated over all the data disregarding the class labels.

$$Precision_{micro} = \frac{\sum_c TP}{\sum_c TP + \sum_c FP} \tag{4}$$

$$Recall_{micro} = \frac{\sum_c TP}{\sum_c TP + \sum_c FN} \tag{5}$$

$$f1_{micro} = \frac{2 * Precision_{micro} * Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{6}$$

However, by this definition for a dataset where the majority of labels belong to one class, the precision score remains high even if the model has not detected the labels of the other class correctly. Therefore, for a better evaluation of the model, we compute the precision, recall, f1 score for each class separately and report the average score on the two classes. This is also known as macro score in the *scikit-learn* python library [15].

$$Precision_{macro} = \frac{1}{|c|} \sum_c \frac{TP}{TP + FP} \tag{7}$$

$$Recall_{macro} = \frac{1}{|c|} \sum_c \frac{TP}{TP + FN} \tag{8}$$

$$f1_{macro} = \frac{2 * Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}} \tag{9}$$

$\psi_{ij}(x_i, x_j)$	$x_j = user$	$x_j = bot$
$x_i = user$	$0.5 + w\epsilon$	$0.5 - w\epsilon$
$x_i = bot$	$0.5 - w\epsilon$	$0.5 + w\epsilon$

Table 3: Edge potentials matrices

Node	$P(user)$	$P(bot)$
User	0.99	0.01
Bot	0.01	0.99
Unknown/Validation	0.5	0.5

Table 4: Node potential based on the original state

5.1 Comparsion with MLP and Belief Propagation

We further evaluated our approach by comparing it with two other methods. As graph convolutional neural networks take both the feature set and the graph structure into consideration, we demonstrate the performance of this method by comparing it with multi layer perceptron (MLP) and belief propagation (BP).

The MLP classifier is trained based on the feature set defined in the Features section. The input layer takes the feature vectors normalized to values between 0 and 1 for each account. The hidden layers consist of two layers with 25 and 10 neurons respectively and use a rectified linear unit as the transfer function. The log loss function is optimized using stochastic gradient descent with a learning rate of 0.0001.

On the other hand, the Belief Propagation algorithm runs solely on the graph structure. The Belief Propagation (BP) algorithm originally proposed by Judea Pearl [14] infers a node's label from some prior knowledge about that node and other neighboring nodes by iteratively passing messages between all pairs of nodes in the graph. The message sent indicates nodes' beliefs regarding the state of their neighbours. For details please refer to [19]. In this experiments we adopted the original BP with the node and edge potential metrics indicated in Table 3 and 4. Furthermore, we ran the experiment with 7 iterations as the messages passed across nodes had no significant changes after 7 iterations.

We plotted the Receiver Operating Characteristic (ROC) for the different models as shown in figure 4. We observe that the area under the ROC curve is 94% for the GCNN approach which is 8% and 16% percent higher than the MLP and BP approach respectively.

While neural networks have shown to perform well in various domains, they are often considered as black boxes when it comes to why they result in such outputs. Interpreting each entry of the output and the meaning of the embedding vectors generated remains open question and topic to investigate for future research.

6 CONCLUSION AND FUTURE WORK

In this paper, we have examined a new approach for detecting malicious accounts and social bots on Twitter by using graph convolutional networks. The main idea of our method is to employ the graph structure and relationships of Twitter accounts for classifying the accounts. Each account aggregates the feature information from its neighborhood. To demonstrate the efficacy of our proposal,

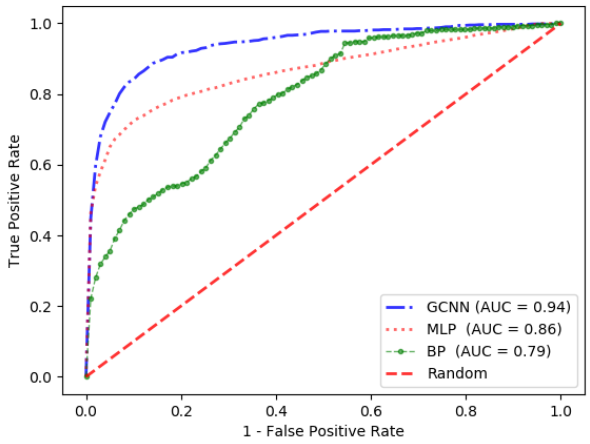


Figure 4: Comparison of the area under curve of different algorithms.

we have worked on a previous well-known dataset in bot detection. Results show that our approach outperforms the state of the art classification algorithms with 8% improvement in the area under curve accuracy.

Since the Twitter API has a limit of 15 requests per rate limit window every 15 minutes, building the Twitter graph structure based on the follower and friend relation of the accounts is not an easy task. We are aware this may be considered as a limitation to our approach. It can thus be suggested to build the graph structure based on the retweet graph of user accounts. Finally, a specific extension for future work is to deploy this method in real time on Twitter's streaming API for spambot detection.

ACKNOWLEDGMENTS

The authors would like to thank the HPI Future SOC Lab for providing access to the resources during the period of Fall 2018.

REFERENCES

[1] N. Chavoshi, H. Hamooni, and A. Mueen. 2016. DeBot: Twitter Bot Detection via Warped Correlation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 817–822. https://doi.org/10.1109/ICDM.2016.0096

[2] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2017. Temporal Patterns in Bot Activities. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1601–1606. https://doi.org/10.1145/3041021.3051114

[3] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. 963–972. https://doi.org/10.1145/3041021.3055135

[4] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. 2016. DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. *IEEE Intelligent Systems* 31, 5 (Sept 2016), 58–64. https://doi.org/10.1109/MIS.2016.29

[5] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. International World Wide Web Conferences Steering Committee, 273–274. https://doi.org/10.1145/2872518.2889302

	$Precision_{macro}$	$Recall_{macro}$	$f1_{macro}$
MLP	0.81	0.73	0.77
BP	0.56	0.54	0.55
GCNN (*with features* 1, 2, 3, 4)	0.85	0.77	0.80
GCNN (*with features* 5, 6)	0.80	0.69	0.72
GCNN (*All features*)	0.89	0.80	0.84

Table 5: Comparison of different algorithms on the dataset

[6] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59, 7 (June 2016), 96–104. https://doi.org/10.1145/2818717

[7] Palash Goyal, Homa Hosseinmardi, Emilio Ferrara, and Aram Galstyan. 2018. Embedding Networks with Edge Attributes. In *Proceedings of the 29th on Hypertext and Social Media (HT '18)*. ACM, New York, NY, USA, 38–42. https://doi.org/10.1145/3209542.3209571

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 1024–1034. http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs.pdf

[9] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *CoRR* abs/1709.05584 (2017). arXiv:1709.05584 http://arxiv.org/abs/1709.05584

[10] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[11] Srijan Kumar and Neil Shah. 2018. False Information on Web and Social Media: A Survey. *CoRR* abs/1804.08559 (2018). arXiv:1804.08559 http://arxiv.org/abs/1804.08559

[12] Kyumin Lee, Brian David Eoff, and James Caverlee. 2011. Seven months with the devils: a long-term study of content polluters on Twitter. In *In AAAI IntâĂŹl Conference on Weblogs and Social Media (ICWSM)*.

[13] Mehwish Nasim, Andrew Nguyen, Nick Lothian, Robert Cope, and Lewis Mitchell. 2018. Real-time Detection of Content Polluters in Partially Observable Twitter Networks. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1331–1339. https://doi.org/10.1145/3184558.3191574

[14] Judea Pearl. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[16] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. AAAI Press.

[17] Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Transactions on Information Forensics and Security* (2013).

[18] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing Spammers' Social Networks for Fun and Profit: A Case Study of Cyber Criminal Ecosystem on Twitter. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 71–80. https://doi.org/10.1145/2187836.2187847

[19] Jonathan S Yedidia, William T Freeman, and Yair Weiss. 2003. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium* 8 (2003), 236–239.

Less is More: Semi-Supervised Causal Inference for Detecting Pathogenic Users in Social Media

Hamidreza Alvari
Arizona State University
halvari@asu.edu

Elham Shaabani
Arizona State University
eshaaban@asu.edu

Soumajyoti Sarkar
Arizona State University
ssarka18@asu.edu

Ghazaleh Beigi
Arizona State University
gbeigi@asu.edu

Paulo Shakarian
Arizona State University
shak@asu.edu

ABSTRACT

Recent years have witnessed a surge of manipulation of public opinion and political events by malicious social media actors. These users are referred to as "Pathogenic Social Media (PSM)" accounts. PSMs are key users in spreading misinformation in social media to viral proportions. These accounts can be either controlled by real users or automated bots. Identification of PSMs is thus of utmost importance for social media authorities. The burden usually falls to automatic approaches that can identify these accounts and protect social media reputation. However, lack of sufficient labeled examples for devising and training sophisticated approaches to combat these accounts is still one of the foremost challenges facing social media firms. In contrast, unlabeled data is abundant and cheap to obtain thanks to massive user-generated data. In this paper, we propose a semi-supervised causal inference PSM detection framework, SemiPsm, to compensate for the lack of labeled data. In particular, the proposed method leverages unlabeled data in the form of manifold regularization and only relies on cascade information. This is in contrast to the existing approaches that use exhaustive feature engineering (e.g., profile information, network structure, etc.). Evidence from empirical experiments on a real-world ISIS-related dataset from Twitter suggests promising results of utilizing unlabeled instances for detecting PSMs.

CCS CONCEPTS

• **Information systems → Social networks**; • **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

Semi-supervised learning; Causal inference; Pathogenic users; Social media

ACM Reference Format:
Hamidreza Alvari, Elham Shaabani, Soumajyoti Sarkar, Ghazaleh Beigi, and Paulo Shakarian . 2019. Less is More: Semi-Supervised Causal Inference for Detecting Pathogenic Users in Social Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316500

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316500

1 INTRODUCTION

Over the past years, social media play major role in massive dissemination of misinformation online. Political events and public opinion on the Web and social networks have been allegedly manipulated by different forms of accounts including real users and automated software (a.k.a social bots or sybil accounts). Pathogenic Social Media (PSM) accounts are among those that are responsible for such a massive spread of disinformation online and swaying normal users' opinion [2, 35]. These accounts (1) are usually owned by either normal users or automated bots, (2) seek to promote or degrade certain ideas; and (3) can appear in many forms such as terrorist supporters (e.g., ISIS supporters), water armies or fake news writers. Understanding the behavior of PSMs will allow social media to take countermeasures against their propaganda at the early stage and reduce their threat to the public.

The problem of identification of PSMs has long been addressed in the past by the research community mostly in the form of bot detection. Several approaches especially supervised learning methods have been proposed in the literature and they have shown promising results [30]. However, for the most part, these approaches are often based on labeled data and exhaustive feature engineering. Examples of such feature groups include but are not limited to content, sentiment of posts, profile information and network features. These approaches are thus very expensive as they require significant amount of efforts to design features and annotate large labeled datasets. In contrast, unlabeled data is ubiquitous and cheap to collect thanks to the massive user-generated data produced on a daily basis. Thus, in this work we set out to examine if unlabeled instances can be utilized to compensate for the lack of enough labeled data.

Present Work. In this paper, semi-supervised causal inference is tailored to detect PSMs who are promoters of misinformation online. We cast the problem of identifying PSMs as an optimization problem and propose a semi-supervised causal learning framework which utilizes unlabeled examples through manifold regularization [11]. In particular, we incorporate causality-based features extracted from users' activity log (i.e., cascades of retweets) as regularization terms into the optimization problem. In this work, causal inference is leveraged in an effort to capture whether or not PSMs exert causal influences while making a message viral. Our

causality-based features are built upon *Suppes' theory of probabilistic causation* [40] whose central concept is *prima facie causes*: an event to be recognized as a cause, must occur before the effect and must lead to an increase of the likelihood of observing the effect. While there exists a prolific literature on causality and their great impact in the computer-science community (see [32] for instance), we build our foundation on *Suppes' theory* as it is computationally less complex.

Key idea and highlights. To summarize, this paper makes the following main contributions:

- We frame the problem of detecting PSM accounts as an optimization problem and present a Laplacian semi-supervised causal inference SemiPsm for solving it. The unlabeled data are utilized via manifold regularization.
- Manifold regularization used in the resultant optimization formulation is built upon causality-based features created on a notion of *Suppes' theory of probabilistic causation*.
- We conduct a suite of experiments using different supervised and semi-supervised methods. Empirical experiments on a real-world ISIS-related dataset from Twitter suggests the effectiveness of the proposed semi-supervised causal inference over the existing methods.

The remainder of the paper is organized as follows. In Section 2, we present the proposed framework. Section 3 summarizes the empirical experiments on an ISIS-related dataset from Twitter. In Section 4 we review the state-of-the-art methods. We conclude the paper in Section 5 by presenting the future work.

2 THE PROPOSED METHOD

In this section, we first provide the causal inference used to extract features out of users' activity log. Then, we detail the proposed semi-supervised causal inference, namely SemiPsm, for detecting PSM accounts.

2.1 Causal Inference

We follow the convention of [23] and assume an *action log* $\mathcal{A}$ of the form *Actions(User,Action,Time)*, which contains tuples (i, a_i, t_i) indicating that user i has performed action a_i at time t_i. For ease of exposition, we slightly abuse the notation and use the tuple (i, m, t) to indicate that user i has posted (tweeted/retweeted) message m at time t. For a given message m we define a *cascade* of actions as $\mathcal{A}_m = \{(i, m', t) \in \mathcal{A} | m' = m\}$. User i is called m-participant if there exists t_i such that $(i, m, t_i) \in \mathcal{A}$. Users who have adopted a message in the early stage of its life span are called *key users*:

Definition 1 (Key Users). *Given message m, m-participant i and cascade $\mathcal{A}_m$, we say user i is a key user iff user i precedes at least ϕ fraction of other m-participants where $\phi \in (0, 1)$. In other words, $|\mathcal{A}_m| \times \phi \leq |\{j | \exists t' : (j, m, t') \in \mathcal{A} \wedge t < t'\}|$, where $|.|$ is the cardinality of a set.*

Next, we shall define viral messages as follows.

Definition 2 (Viral Messages). *Given a threshold θ, we say a message $m \in \mathbf{M}$ is viral iff $|\mathcal{A}_m| \geq \theta$. We denote a set of all viral messages by $\mathbf{M}_{vir}$.*

The prior probability of a message going viral is $\rho = |\mathbf{M}_{vir}|/|\mathbf{M}|$. The probability of a message going viral given key user i has participated in, is computed as follows:

$$\rho_i = \frac{|\{m | m \in \mathbf{M}_{vir} \wedge i \text{ is a key user}\}|}{|\{m | m \in \mathbf{M} \wedge i \text{ is a key user}\}|} \tag{1}$$

The probability that key users i and j tweet/retweet message m chronologically and make it viral, is computed as:

$$p_{i,j} = \frac{|\{m \in \mathbf{M}_{vir} | \exists t, t' : t < t' \wedge (i, m, t), (j, m, t') \in \mathcal{A}\}|}{|\{m \in \mathbf{M} | \exists t, t' : t < t' \wedge (i, m, t), (j, m, t') \in \mathcal{A}\}|} \tag{2}$$

To examine how causal user i was in helping a message m going viral, we shall explore what will happen if we exclude user i from m. We define the probability that *only* key user j has made a message m viral, i.e. user i has not posted m or does not precede j as:

$$p_{\neg i,j} = \frac{|\{m \in \mathbf{M}_{vir} | \exists t' : (j, m, t') \in \mathcal{A} \wedge \nexists t : t < t', (i, m, t) \in \mathcal{A}\}|}{|\{m \in \mathbf{M} | \exists t' : (j, m, t') \in \mathcal{A} \wedge \nexists t : t < t', (i, m, t) \in \mathcal{A}\}|} \tag{3}$$

In this work we adopt the notion of *prima facie causes* which is at the core of Suppes' theory of probabilistic causation [40] and introduce causality metrics. According to this theory, *a certain event to be recognized as a cause, must occur before the effect and must lead to an increase of the likelihood of observing the effect.* Accordingly, prima facie causal users are defined as follows:

Definition 3 (Prima Facie Causal Users). *A user i is said to be Prima Facie causal user for cascade $\mathcal{A}_m$ iff: (1) user i is a key user of m, (2) $m \in \mathbf{M}_{vir}$, and (3) $\rho_i > \rho$.*

We use the concept of *related users* from a rule-based system [38] which was an extension to the causal inference framework in [28]. Accordingly, we call users i and j m-related if (1) they are Prima Facie causal users for m, and (2) i precedes j. We then define a set of user i's related users as $\mathbf{R}(i) = \{j | j \neq i \text{ and } i, j \text{ are } m\text{-related}\}$.

In this work, we use the time-decay causal metrics introduced in [2] which are built on Suppes' theory. The first metric used in this work is $\mathcal{E}_{K\&M}$ which is computed over a given time interval I as follows:

$$\mathcal{E}^I_{K\&M}(i) = \frac{\sum_{j \in \mathcal{R}(i)} (\mathcal{P}_{i,j} - \mathcal{P}_{\neg i,j})}{|\mathcal{R}(i)|} \tag{4}$$

where $\mathcal{R}(i)$, $\mathcal{P}_{i,j}$, and $\mathcal{P}_{\neg i,j}$ are now defined over I. This metric estimates the causality score of user i in making a message *viral*, by taking the average of $\mathcal{P}_{i,j} - \mathcal{P}_{\neg i,j}$ over $\mathbf{R}(i)$. The intuition here is that user i is more likely to be a cause of message m to become viral than user j, if $\mathcal{P}_{i,j} - \mathcal{P}_{\neg i,j} > 0$. This metric cannot spot all PSMs, hence another metric is defined, namely relative likelihood causality $\mathcal{E}_{rel}$. This metric works by assessing relative difference between $\mathcal{P}_{i,j}$, and $\mathcal{P}_{\neg i,j}$:

$$\mathcal{E}^I_{rel}(i) = \frac{\mathcal{S}(i,j)}{|\mathcal{R}(i)|} \tag{5}$$

where $\mathcal{S}(i,j)$ is defined as follows and α is infinitesimal:

$$\mathcal{S}(i,j) = \begin{cases} \frac{\mathcal{P}_{i,j}}{\mathcal{P}_{\neg i,j} + \alpha} - 1, & \mathcal{P}_{i,j} > \mathcal{P}_{\neg i,j} \\ 1 - \frac{\mathcal{P}_{\neg i,j}}{\mathcal{P}_{i,j}}, & \mathcal{P}_{i,j} \leq \mathcal{P}_{\neg i,j} \end{cases} \tag{6}$$

Two other neighborhood-based metrics were also defined in [2], first of which is computed as:

$$\mathcal{E}^I_{nb}(j) = \frac{\sum_{i \in Q(j)} \mathcal{E}^I_{K\&M}(i)}{|Q(j)|} \tag{7}$$

where $Q(j) = \{i | j \in \mathcal{R}(i)\}$ is the set of all users that user j belongs to their related users sets. Similarly, the second metric is the weighted version of the above metric and is called weighted neighborhood-based causality and is calculated as:

$$\mathcal{E}^I_{wnb}(j) = \frac{\sum_{i \in Q(j)} w_i \times \mathcal{E}^I_{K\&M}(i)}{\sum_{i \in Q(j)} w_i} \tag{8}$$

The aim of this metric is to capture different impacts that users in $Q(j)$ might have on user j.

2.2 Final set of Features

Finally, the causal metrics discussed in the previous section will be fed as features to the semi-supervised framework– this will be described in the next section. The final set of features is in the following generic form ξ^I_k where $k \in \{K\&M, rel, nb, wnb\}$ [2]:

$$\xi^I_k(i) = \frac{1}{|\mathcal{T}|} \sum_{t' \in \mathcal{T}} e^{-\sigma(t-t')} \times \mathcal{E}^{\Delta}_k(i) \tag{9}$$

Here, σ is a scaling parameter of the exponential decay function, $\mathcal{T} = \{t' | t' = t_0 + j \times \delta, j \in \mathbb{N} \wedge t' \leq t - \delta\}$ is a sequence of sliding-time windows, and δ is a small fixed amount of time, which is used as the length of each sliding-time window $\Delta = [t' - \delta, t']$.

In essence, this metric assigns different weights to different time points of a given time interval, inversely proportional to their distance from t (i.e., smaller distance is associated with higher weight). Specifically, it performs the following: it (1) breaks down the given time interval into shorter time periods using a sliding time window, (2) deploys an exponential decay function of the form $f(x) = e^{-\alpha x}$ to account for the time-decay effect, and (3) takes average of the causality values computed over each sliding time window [2].

2.3 Semi-Supervised Causal Inference

Having defined the causality-based features, we now proceed to present the proposed semi-supervised Laplacian SVM framework, SemiPsm. For the rest of the discussion, we shall assume a set of l labeled pairs $\{(x_i, y_i)\}_{i=1}^{l}$ and an unlabeled set of u instances $\{x_{l+i}\}_{i=1}^{u}$, where $x_i \in \mathbb{R}^n$ denotes the causality vector $\xi^I_k(i)$ of user i and $y_i \in \{+1, -1\}$ (PSM or not).

Recall for the standard soft-margin support vector machines, the following optimization problem is solved:

$$\min_{f_\theta \in \mathcal{H}_k} \gamma ||f_\theta||^2_k + C_l \sum_{i=1}^{l} H_1(y_i f_\theta(x_i)) \tag{10}$$

In the above equation, $f_\theta(\cdot)$ is a decision function of the form $f_\theta(\cdot) = w.\Phi(\cdot) + b$ where $\theta = (w, b)$ are the parameters of the model, and $\Phi(\cdot)$ is the feature map which is usually implemented using the kernel trick [17]. Also, the function $H_1(\cdot) = \max(0, 1 - \cdot)$ is the Hinge Loss function. The classical Representer theorem [10] suggests that solution to the optimization problem exists in a Hilbert space $\mathcal{H}_k$ and is of the form $f^*_\theta(x) = \sum_{i=1}^{l} \alpha^*_i \mathbf{K}(x, x_i)$. Here, $\mathbf{K}$ is the $l \times l$ Gram matrix over labeled samples. Equivalently, the above problem can be written as:

$$\min_{w,b,\epsilon} \frac{1}{2}||w||^2_2 + C_l \sum_{i=1}^{l} \epsilon_i \tag{11}$$

$$s.t.\ y_i(w.\Phi(x_i) + b) \geq 1 - \epsilon_i,\ i = 1, ..., l$$

$$\epsilon_i \geq 0,\ i = 1, ..., l \tag{12}$$

Next, we will use the above optimization equation as our basis to derive the formulations for our proposed semi-supervised learner.

The basic assumption behind semi-supervised learning methods is to leverage unlabeled instances in order to restructure hypotheses during the learning process [4]. Here, exogenous information extracted from causality-based features of users is exploited to make a better use of the unlabeled examples. To do so, we first introduce matrix $\mathbf{F}$ over both of the labeled and unlabeled samples with $\mathbf{F}_{ij} = ||\Phi(x_i) - \Phi(x_j)||_2$ in $||.||_2$ norm. This way, we force instances x_i and x_j in our dataset to be relatively 'close' to each other [8], i.e., having a same label, if their corresponding causal-based feature vectors are close. To account for this, a regularization term is added to the standard equation and the following optimization is solved:

$$\min_{f_\theta \in \mathcal{H}_k} \frac{1}{2} \sum_{i=1}^{l} \mathbf{F}_{ij} ||f_\theta(x_i) - f_\theta(x_j)||^2_2 = \mathbf{f}^T_\theta \mathcal{L}^T \mathbf{f}_\theta \tag{13}$$

where $\mathbf{f} = [f(x_1), ..., f(x_{l+u})]^T$ and $\mathcal{L}$ is the Laplacian matrix based on $\mathbf{F}$ given by $\mathcal{L} = \mathbf{D} - \mathbf{F}$, and $\mathbf{D}_{ii} = \sum_{j=1}^{l+u} \mathbf{F}_{ij}$. The intuition here is that causal pairs are more likely to have same labels than others.

Following the notations used in [11] and by including our regularization term, we would extend the standard equation by solving the following optimization:

$$\min_{f_\theta \in \mathcal{H}_k} \gamma ||f_\theta||^2_k + C_l \sum_{i=1}^{l} H_1(y_i f_\theta(x_i)) + C_r \mathbf{f}^T_\theta \mathcal{L} \mathbf{f}_\theta \tag{14}$$

Again, solution in $\mathcal{H}_k$ would be in the following form $f^*_\theta(x) = \sum_{i=1}^{l+u} \alpha^*_i \mathbf{K}(x, x_i)$. Here $\mathbf{K}$ is the $(l+u) \times (l+u)$ Gram matrix over all samples. The Eq. 14 could be then written as follows:

$$\min_{\alpha,b,\epsilon} \frac{1}{2} \alpha^T \mathbf{K} \alpha + C_l \sum_{i=1}^{l} \epsilon_i + \frac{C_r}{2} \alpha^T \mathbf{K} \mathcal{L} \mathbf{K} \alpha \tag{15}$$

$$s.t.\ y_i(\sum_{j=1}^{l+u} \alpha_j \mathbf{K}(x_i, x_j) + b) \geq 1 - \epsilon_i,\ i = 1, ..., l$$

$$\epsilon_i \geq 0,\ i = 1, ..., l \tag{16}$$

With introduction of the Lagrangian multipliers β and γ, we write the Lagrangian function of the above equation as follows:

$$L(\alpha, \epsilon, b, \beta, \gamma) = \frac{1}{2} \alpha^T \mathbf{K}(I + C_r \mathcal{L})\alpha + C_l \sum_{i=1}^{l} \epsilon_i \tag{17}$$

$$- \sum_{i=1}^{l} \beta_i (y_i (\sum_{j=1}^{l+u} \alpha_j \mathbf{K}(x_i, x_j) + b) - 1 + \epsilon_i) - \sum_{i=1}^{l} \gamma_i \epsilon_i$$

Obtaining the dual representation, requires taking the following steps:

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} \beta_i y_i = 0 \tag{18}$$

$$\frac{\partial L}{\partial \epsilon_i} = 0 \rightarrow C_l - \beta_i - \gamma_i = 0 \rightarrow 0 \leq \beta_i \leq C_l \tag{19}$$

With the above equations, we formulate the reduced Lagrangian as a function of only α and β as follows:

$$L^R(\alpha, \beta) = \frac{1}{2}\alpha^T \mathbf{K}(I + C_r\mathcal{L})\alpha - \alpha^T \mathbf{K}\mathbf{J}^T \mathbf{Y}\beta + \sum_{i=1}^{l} \beta_i \tag{20}$$

In the above equation, $\mathbf{J} = [\mathbf{I}\ \mathbf{0}]$ is a $l \times (l+u)$ matrix, $\mathbf{I}$ is the $l \times l$ identity matrix and Y is a diagonal matrix consisting of the labels of the labeled examples. We first take the derivative of L^R with respect to α and then set $\frac{\partial L^R(\alpha,\beta)}{\partial \alpha} = 0$. We have the following equation:

$$\mathbf{K}(I + C_r\mathcal{L})\alpha - \mathbf{K}\mathbf{J}^T \mathbf{Y}\beta = 0 \tag{21}$$

Accordingly, we obtain α^* by solving the following equation:

$$\alpha^* = (I + C_r\mathcal{L})^{-1}\mathbf{J}^T \mathbf{Y}\beta^* \tag{22}$$

Next, we obtain the dual problem in the form of a quadratic programming problem by substituting α back in the reduced Lagrangian function Eq. 20:

$$\beta^* = \text{argmax}_{\beta \in \mathbb{R}^l} \ -\frac{1}{2}\beta^T \mathbf{Q}\beta + \sum_{i=1}^{l} \beta_i \tag{23}$$

$$s.t. \ \sum_{i=1}^{l} \beta_i y_i = 0$$

$$0 \leq \beta_i \leq C_l \tag{24}$$

where $\beta = [\beta_1, ..., \beta_l]^T \in \mathbb{R}^l$ are the Lagrangian multipliers and $\mathbf{Q}$ is obtained as follows:

$$\mathbf{Q} = \mathbf{Y}\mathbf{J}\mathbf{K}(I + (C_r\mathcal{L})\mathbf{K})^{-1}\mathbf{J}^T\mathbf{Y} \tag{25}$$

We summarize the proposed semi-supervised framework in Algorithm 1. Our optimization problem is very similar to the standard optimization problem solved for SVMs, hence we use a standard optimizer for SVMs to solve our problem.

2.4 Computational Complexity

Here, we will explain the scalability of the algorithm in terms of big-O notation for both constituents of the proposed framework separately. For the first part of the approach, given a set of $\mathcal{A}$ cascades, and average number of $avg(\tau)$ users' actions (i.e., timestamps) in each cascade where $\tau \in \mathcal{A}$, the complexity of computing causality scores is $O(|\mathcal{A}|.(avg(\tau))^2)$ (note on average there are $(avg(\tau))^2$ pairs of users in each cascade). For the second part, i.e., learning the semi-supervised algorithm, the most time-consuming part is calculating the inverse of a dense Gram matrix which leads to $O((l+u)^3)$ complexity, where l and u are number of labeled and unlabeled instances [11].

Algorithm 1 Semi-Supervised Causal Inference for PSM detection (SemiPsm)

Input: $\{(x_i, y_i)\}_{i=1}^{l}$, $\{x_{l+i}\}_{i=1}^{u}$, $\mathcal{F}_1$, $\mathcal{F}_2$, C_l, C_r.
Output: Estimated function $f_\theta : \mathbb{R}^n \rightarrow \mathbb{R}$

1: Construct matrix **F** based on the causality-based features
2: Compute the corresponding Laplacian matrix $\mathcal{L}$.
3: Construct the Gram matrix over all examples using $\mathbf{K}_{ij} = k(x_i, x_j)$ where k is a kernel function.
4: Compute α^* and β^* using Eq. 22 and Eq. 23 and a standard QP solvers.
5: Compute function $f_\theta^*(x) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(x, x_i)$

3 EXPERIMENTS

In this section we conduct experiments on a Twitter ISIS-related dataset and present results for several supervised and semi-supervised approaches. We first explain the dataset and provide some data analysis. Then, we will present the baseline methods. Finally, results and discussion are provided.

3.1 ISIS Twitter Dataset

We collect a dataset (Table 1) of 53 M ISIS related tweets/retweets in Arabic, from Feb 22, 2016 to May 27, 2016. The dataset has different fields including user ID, retweet ID, hashtags, content, posting time. The tweets were collected using 290 different hashtags such as #Terrorism and #StateOfTheIslamicCaliphate. We use a subset of this dataset which contains 35 K cascades of different sizes and durations. There are ~11 M tweets/retweets associated with the cascades. After pre-processing and removing duplicate users from cascades, cascades sizes (i.e. number of associated postings) vary between 20 to 9,571 and take from 10 seconds to 95 days to finish. The log-log distribution of cascades vs. cascade size and the cumulative distribution of duration of cascades are depicted in Figure 1.

Based on the content of tweets in our dataset, PSMs are terrorism-supporting accounts who have participated in viral cascades. We chose to use threshold $\theta = 100$ and take about 6 K viral cascades with at least 100 tweets/retweets. We demonstrate in Figure 2, the total number of users in each cascade suspended by Twitter. We note that he dataset does not have any underlying network. We only focus on the non-textual information in the form of an *action log*. We set $\phi = 0.5$ to select *key users*, i.e., we are looking for the users that participate in the cascades before the number of participants gets twice. After the data collection, we follow [41] and check through Twitter API whether users have been suspended (PSM) or are still active (normal). According to Table 1, less than 24% of the users in our dataset are PSM and the rests are normal.

3.2 Baseline Methods

We compare the proposed method SemiPsm against the following baseline methods. Note for all methods, we only report results when their best settings are used.

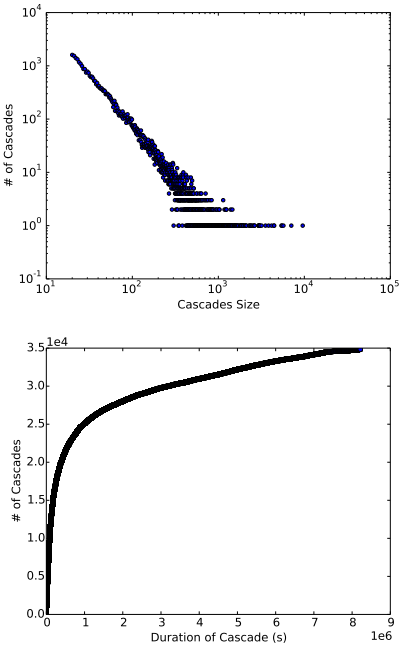


Figure 1: (Top) Log-log distribution of cascades vs. cascade size. (Bottom) Cumulative distribution of duration of cascades.

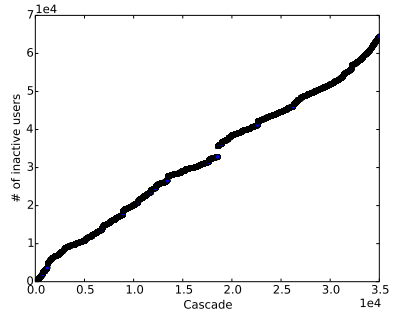


Figure 2: Total inactive users in each cascade.

Table 1: Description of the dataset.

Name	Value	
# of Cascades	35 K	
# of Viral Cascades	6,602	
# of Tweets/Retweets	10,823,168	
# of Users	PSM	Normal
	19,859	65,417

- **LabelSpreading (Rbf Kernel) [44].** This is a graph inference-based label spreading approach with radial basis function (RBF) kernel.
- **Label Spreading (Knn Kernel) [44].** Similar to the previous approach with K-nearest neighbor (KNN) kernel.
- **Lstm [30].** The word-level LSTM approach here is similar to the deep neural network models used for sequential word predictions. We adapt the neural network to a sequence classification problem where the inputs are the vector of words in each tweet and the output is the predicted label of the tweet. We first use the word2vec [31] embedding pre-trained from a set of tweets similar to the data representation in our Twitter dataset.
- **Account-Level (Rf Classifier) [30]** This approach uses the following features of the user profiles: *Statuses Count, Followers Count, Friends Count, Favorites Count, Listed Count, Default Profile, Geo Enables, Profile Uses Background Image, Verified, Protected.* We chose this method over Botometer [42] as it achieved comparable results with far less number of features ([42] uses over 1,500 features)(see also [21]). According to [30], we report the best results when Random Forest (RF) is used.
- **Tweet-Level (Rf Classifier) [30].** Similar to the previous baseline, this method uses only a handful of features extracted from tweets: *retweet count, reply count, favorite count, number of hashtags, number of URLs, number of mentions.* Likewise, we use RF as the classification algorithm.
- **SentiMetrix [39].** This approach was proposed by the top-ranked team in the DARPA Twitter Bot Challenge. We consider all features that we could extract from our dataset. Our features include tweet syntax (average number of hashtags, average number of user mentions, average number of links, average number of special characters), tweet semantics (LDA topics), and user behaviour (tweet spread, tweet frequency, tweet repeats). The proposed approach starts with a small seed set and propagates the labels. Since we have enough labeled data for the training part, we use Random Forest as the learning approach.
- **C2dc [2].** This approach uses time-decay causal community detection-based classification to detect PSM accounts [2]. For community detection, this approach uses Louvain algorithm.

3.3 Results and Discussion

All experiments were implemented in Python 2.7x and run on a machine equipped with an Intel(R) Xeon(R) CPU of 3.50 GHz with 200 GB of RAM running Linux. The proposed approach was implemented using CVXOPT[1] package. Furthermore, we split the whole dataset into 50% training and 50% test sets for all experiments. We report results in terms of F1-score in tables 2 and 3. For any approach that requires special tuning of parameters, we conducted grid search to choose the best set of parameters. Specifically, for the proposed approach, we set the penalty parameter as $C_l = 0.6$ and the regularization parameter $C_r = 0.2$, and used linear kernel. For LabelSpreading (Rbf), the default vale of $\gamma = 20$ was used and for LabelSpreading (Knn), number of neighbors was set to 5. Also, for random forest we used 200 estimators and the 'entropy' criterion was used. For computing k nearest neighbors in C2dc, we set $k = 10$.

[1] http://cvxopt.org/

Table 2: F1-score results of various methods on the labeled data. For semi-supervised learners, the size of the unlabeled data is fixed to 10% of the training set. The best performance is in bold.

Learner	F1-score
SemiPsm (Causal Features)	**0.94**
SemiPsm (Account-Level Features)	0.89
SemiPsm (Tweet-Level Features)	0.88
LabelSpreading (Knn/Causal Features)	0.89
LabelSpreading (Rbf/Causal Features)	0.88
Account-Level (Rf Classifier)	0.88
Tweet-Level (Rf Classifier)	0.82
SentiMetrix	0.54
Lstm	0.41
C2dc	0.4

Table 3: F1-score results of the semi-supervised approaches when causality-based features are used. Results are reported on different portions of the unlabeled data. The best performance is in bold.

	Percentage of Unlabeled Data				
	10%	**20%**	**30%**	**40%**	**50%**
SemiPsm	**0.94**	**0.93**	**0.91**	**0.9**	**0.88**
LabelSpreading (Knn)	0.89	0.88	0.87	0.85	0.81
LabelSpreading (Rbf)	0.88	0.86	0.85	0.82	0.80

Furthermore for LSTM, we preprocessed the individual tweets in line with the steps mentioned in [37]. Since the content of the tweets are in Arabic, we replaced special characters that were present in the text with their Arabic counterparts if they were present. We used word vectors of dimensions 100 and deployed the skip-gram technique for obtaining the word vectors where the input is the target word, while the outputs are the words surrounding the target words. To model the tweet content in a manner that uses it to predict whether an account is PSM or not, we used Long Short Term Memory (LSTM) models [25]. For the LSTM architecture, we used the first 20 words in the tokenized Arabic text of each tweet and use padding in situations where the number of tokens in a tweet are less than 20. We used 30 units in the LSTM architecture (many to one). The output of the LSTM layer was fed to a dense layer of 32 units with ReLU activations. We added dropout regularization following this layer to avoid overfitting and the output was then fed to a dense layer which outputs the category of the tweets.

We depict in Table 2 classification performance of all approaches on the labeled data. For the proposed framework SemiPsm, we examine three sets of features (1) causality-based features, (2) account-level features [30]; and (3) tweet-level features [30]. For the graph inference-based semi-supervised algorithms, i.e., LabelSpreading (Rbf) and LabelSpreading (Knn), we only report results where causality-based features are used as they achieved best performance with them. As it is observed from the table, the best results in terms of F1-score belong to SemiPsm where causality-based features are used. The runner-up is SemiPsm with account-level features and the next best approach is SemiPsm where tweet-level features are deployed. This clearly demonstrates the significance of using manifold regularization in the Laplacian semi-supervised framework over using other semi-supervised methods, LabelSpreading (Rbf) and LabelSpreading (Knn).

We further note that the supervised classifier Random Forest using both of the account-level and tweet-level features and the whole labeled dataset achieve worse or comparable results to the semi-supervised learners. The fact that obtaining several tweet and account-level features is not trivial and do not necessarily lead to the best classification performance, motivates us to use semi-supervised algorithms which use less number of labeled examples, and yet achieve competing performance. We also obtain an F1-score of 0.41 when LSTM is used– the poor performance of the this neural network model can be attributed to the raw Arabic text content. It suggests that the Arabic tokens as a representation might not be very informative about the category of accounts it has been generated from and some kind of weighting might be necessary before the LSTM module is used.

Also, Table 3 shows the classification performance of the semi-supervised approaches with causality-based features. The results are achieved using different portions of the unlabeled data, i.e., {10%, 20%, 30%, 40%, 50%} of the training set. As it is seen in the table, SemiPsm achieves the best performance on different portions of the unlabeled data compared to the other semi-supervised learners, while performances of all approaches deteriorate with increasing the percentage of the unlabeled data. Furthermore, SemiPsm still outperforms all other supervised methods as well as Lstm and C2dc when up to 50% of the data has been made unlabeled.

Observations. Overall, this paper makes the following observations:

- Among the semi-supervised learners used in this study, SemiPsm achieves the best classification performance suggesting the significance of using unlabeled instances in the form of manifold regularization. Manifold regularization is shown effective in boosting the classification performance, with three different sets of features confirming this.
- Causality-based features achieve the best performance via both Laplacian and graph inference-based semi-supervised settings. This lies at the inherent property of the causality-based features– they are designed to show whether or not user i exerts a causal influence on j. This is effective in capturing PSMs as they are key users in making a message viral.
- Compared to the supervised methods Account-Level (Rf) and Tweet-Level (Rf), semi-supervised learners achieve either comparable or best results, suggesting promising results with less number of labeled examples.
- Among the supervised methods Account-Level (Rf) and Tweet-Level (Rf), the former achieves higher F1-score indicating that account-level features are more useful in boosting the performance, although they are harder to obtain [30].
- Semi-supervised learners achieve best or comparable results with supervised learners, even with up to 50% of the data made unlabeled. This clearly shows the superiority of using unlabeled examples over labeled ones.

4 RELATED WORK

The explosive growth of the Web has raised numerous security and privacy issues. Mitigating these concerns has been studied from several aspects [1, 3, 5–7, 9, 13, 14, 18]. Our work is related to a number of research directions. Below, we will summarize some of the state-of-the-art methods in each category while highlighting their differences with our work.

Identifying PSM accounts. Compared to [34] which uses causal inference to detect PSM accounts, our work utilizes time-decay causal inference (using sliding-time window) which allows for early detection of PSM. In contrast to [2] where a causal community detection algorithm is proposed to leverage communities of PSM accounts in order to achieve higher performance, our work proposes a semi-supervised causal inference algorithm that achieves reasonable performance using less labeled data by utilizing unlabeled data.

Social Spam/Bot Detection. Recently, DARPA organized a Twitter bot challenge to detect "influence bots" [39]. Among the participants, the work of [14], used similarity to cluster accounts and uncover groups of malicious users. The work of [42] presented a supervised framework for bot detection which uses more than thousands features. In a different attempt, the work of [24] studied the problem of spam detection in Wikipedia using different spammers behavioral features. There also exist some studies in the literature that have addressed (1) differences between humans and bots [16], (2) different natures of bots [42] or (3) differences between bots and human trolls [13]. For example the work of [16] conducted a series of measurements in order to distinguish humans from bots and cyborgs, in term of tweeting behavior, content, and account properties. To do so, they used more than 40 million tweets posted by over 500 K users. Then, they performed analysis and find groups of features that are useful for classifying users into human, bots and cyborgs. They concluded that entropy and certain account properties can be very helpful in differentiating between those accounts. In a different attempt, some other studies have tried to differentiate between several natures of bots. For instance, in the work of [42], authors performed clustring analysis and revealed specific behavioral groups of accounts. Specifically, they identified different types of bots such as *spammers*, *self promoters*, and *accounts that post content from connected applications*, using manual investigation of samples extracted from clusters. Their cluster analysis emphasized that Twitter hosts a variety of users with diverse behaviors; that is in some cases the boundary between human and bot users is not sharp, i.e. some account exhibit characteristics of both.

Also, the work of [13], uses Twitter data to quantify the impact of Russian trolls and bots on amplifying polarizing and anti-vaccine tweets. They first used the Botometer API to assign bot probabilities to the users in the dataset and divided the whole dataset into 3 categories: those with scores less than 20% (very likely to be human), between 20% and 80% (e.g., cyborgs with uncertain provenance) and above 80% (high likely to be bots). Then, they posed two research questions: (1) Are bots and trolls more likely to tweet about vaccines?, and (2) Are bots and trolls more likely to tweet polarizing and anti-vaccine content? Their analysis demonstrated that Twitter bots and trolls significantly impact on online discussion about vaccination and this differs by account type. For example, Russian trolls and bots post content about vaccination at higher rates compared to an average user. Also, according to this study, troll accounts and content polluters (e.g., dissemination of malware, unsolicited commercial content, etc.) post anti-vaccine tweets 75% more than average users. In contrast, spambots which can be easily distinguished from humans, are less likely to promote anti-vaccine messages. Their closing remarks suggest strongly that distinguishing between malicious actors (bots, trolls, cyborgs, and human users) is difficult and thus anti-vaccine messages may be disseminated at higher rates by a combination of these malicious actors.

In contrast to the above works, our work does not deploy any extra information (e.g., user-related attributes or network-based features) other than users' actions (i.e., cascade with timestamps). It is also worthwhile to note that most of the existing well-known bot detection algorithms such as Botometer [19] leverage over one thousand features in order to detect high-likely bots.

Fake News Identification. A growing body of research is addressing the impact of bots in manipulating political discussion, including the 2016 U.S. presidential election [36] and the 2017 French election [20]. For example, [36] analyzes tweets following recent U.S. presidential election and found evidences that bots played key roles in spreading fake news.

Identifying Instigators. There are some work on instigator detection [22, 33] and outbreak prediction [18]. In [29], authors performed classification to detect users who adopt popular items. In [45], authors designed an approach for information source detection and in particular initiator of a cascade. Our work is focused on a set of users who *might* or *might not* be initiators. Our work is different from these works since we leverage causality analysis to detect causes of popularity of messages that go viral.

Extremism and Water Armies Detection. The work of [26] designed a behavioral model to detect extremists. Authors in [12] performed iterative vertex clustering and classification to identify Islamic Jihadists on Twitter. The works of [15, 43] also used user behavioral and domain-specific attributes to detect water armies. Our work also differs from these works as we do not use any features such as network/user attributes.

Causal Reasoning. As opposed to [27, 28, 38] which deal with preconditions as single atomic propositions, we use rules with preconditions of more than one atomic propositions.

5 CONCLUSION

We presented a semi-supervised Laplacian SVM to detect PSM users in social media who are promoters of misinformation spread. We cast the problem of identifying PSMs as an optimization problem and introduced a Laplacian semi-supervised SVM via utilizing unlabeled examples through manifold regularization. We examined different sets of features extracted from users activity log (in the form of cascades of retweets) as regularization terms: (1) causality-based features; and (2) LSTM-based features. Our causality-based features were built upon *Suppes' theory of probabilistic causation*. The LSTM-based features were extracted via LSTM which has shown promising results for different tasks in the literature.

In future, we would like to replicate the study by feeding other sets of features such as time-series features and those extracted

using LSTM to the semi-supervised framework. Also, we plan to investigate other forms of causality inferences and other regularization terms to seek if we can further improve the classification performance by distinguishing between different types of PSMs.

6 ACKNOWLEDGMENT

Some of the authors are supported through the DoS and DoD Minerva program.

REFERENCES

[1] Hamidreza Alvari, Soumajyoti Sarkar, and Paulo Shakarian. 2019. Detection of Violent Extremists in Social Media. *IEEE Conference on Data Intelligence and Security* (2019).

[2] Hamidreza Alvari, Elham Shaabani, and Paulo Shakarian. 2018. Early Identification of Pathogenic Social Media Accounts. *IEEE Intelligent and Security Informatics* (2018).

[3] Hamidreza Alvari, Paulo Shakarian, and JE Kelly Snyder. 2016. A non-parametric learning approach to identify online human trafficking. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. IEEE, 133–138.

[4] Hamidreza Alvari, Paulo Shakarian, and JE Kelly Snyder. 2017. Semi-supervised learning for detecting human trafficking. *Security Informatics* 6, 1 (2017), 1.

[5] Ghazaleh Beigi, Ruocheng Guo, Alexander Nou, Yanchao Zhang, and Huan Liu. 2019. Protecting User Privacy: An Approach for Untraceable Web Browsing History and Unambiguous User Profiles. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 213–221.

[6] Ghazaleh Beigi, Mahdi Jalili, Hamidreza Alvari, and Gita Sukthankar. 2014. Leveraging community detection for accurate trust prediction. (2014).

[7] Ghazaleh Beigi and Huan Liu. 2018. Privacy in social media: Identification, mitigation and applications. *arXiv preprint arXiv:1808.02191* (2018).

[8] Ghazaleh Beigi and Huan Liu. 2018. Similar but different: Exploiting usersâĂŹ congruity for recommendation systems. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 129–140.

[9] Ghazaleh Beigi, Kai Shu, Yanchao Zhang, and Huan Liu. 2018. Securing Social Media User Data-An Adversarial Approach. *Proceedings of the 29th on Hypertext and Social Media* (2018), 156–173.

[10] Misha Belkin, Partha Niyogi, and Vikas Sindhwani. 2005. On manifold regularization.. In *AISTATS*. Citeseer.

[11] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research* 7, Nov (2006), 2399–2434.

[12] Matthew C Benigni, Kenneth Joseph, and Kathleen M Carley. 2017. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PloS one* (2017).

[13] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American journal of public health* 108, 10 (2018), 1378–1384.

[14] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering Large Groups of Active Malicious Accounts in Online Social Networks. In *CCS*.

[15] Cheng Chen, Kui Wu, Srinivasan Venkatesh, and Xudong Zhang. 2011. Battling the Internet Water Army: Detection of Hidden Paid Posters. *CoRR* (2011).

[16] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing* 9, 6 (2012), 811–824.

[17] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[18] Peng Cui, Shifei Jin, Linyun Yu, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Cascading Outbreak Prediction in Networks: A Data-driven Approach. In *KDD*.

[19] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. International World Wide Web Conferences Steering Committee.

[20] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. (2017).

[21] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* (2016).

[22] Huang Chung-Yuan Fu, Yu-Hsiang and Chuen-Tsai Sun. 2015. Identifying Super-Spreader Nodes in Complex Networks. *Mathematical Problems in Engineering* (2015).

[23] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. 2010. Learning Influence Probabilities in Social Networks. In *WSDM*.

[24] Thomas Green and Francesca Spezzano. 2017. Spam Users Identification in Wikipedia Via Editing Behavior. *ICWSM* (2017).

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[26] Jytte Klausen, Christopher Marks, and Tauhid Zaman. 2016. Finding Online Extremists in Social Networks. *CoRR* abs/1610.06242 (2016).

[27] Samantha Kleinberg. 2011. A Logic for Causal Inference in Time Series with Discrete and Continuous Variables. In *IJCAI*.

[28] Samantha Kleinberg and Bud Mishra. 2009. The temporal logic of causal structures. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 303–312.

[29] Takuya Konishi, Tomoharu Iwata, Kohei Hayashi, and Ken-Ichi Kawarabayashi. 2016. Identifying Key Observers to Find Popular Information in Advance. In *IJCAI*.

[30] Sneha Kudugunta and Emilio Ferrara. 2018. Deep Neural Networks for Bot Detection. *arXiv preprint arXiv:1802.04289* (2018).

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[32] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York, NY, USA.

[33] Sen Pei, Lev Muchnik, JosÃľ S. Andrade Jr., Zhiming Zheng, and HernÃąn A. Makse. 2014. Searching for superspreaders of information in real-world social media. *CoRR* (2014).

[34] E. Shaabani, R. Guo, and P. Shakarian. 2018. Detecting Pathogenic Social Media Accounts without Content or Network Structure. In *IEEE Conference on Data Intelligence and Security*.

[35] Elham Shaabani, Ashkan Sadeghi-Mobarakeh, Hamidreza Alvari, and Paulo Shakarian. 2019. An End-to-End Framework to Identify Pathogenic Social Media Accounts on Twitter. *IEEE Conference on Data Intelligence and Security* (2019).

[36] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. (2017).

[37] Abu Bakr Soliman, Kareem Eissa, and Samhaa R El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science* 117 (2017), 256–265.

[38] Andrew Stanton, Amanda Thart, Ashish Jain, Priyank Vyas, Arpan Chatterjee, and Paulo Shakarian. 2015. Mining for Causal Relationships: A Data-Driven Study of the Islamic State. *CoRR* (2015).

[39] V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer. 2016. The DARPA Twitter Bot Challenge. (2016).

[40] Patrick Suppes. 1970. A Probabilistic Theory of Causality. (1970).

[41] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *ACM SIGCOMM conference on Internet measurement conference*.

[42] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *ICWSM* (2017).

[43] Kun Wang, Yang Xiao, and Zhen Xiao. 2014. Detection of internet water army in social network.

[44] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *Advances in neural information processing systems*. 321–328.

[45] Kai Zhu and Lei Ying. 2016. Information Source Detection in the SIR Model: A Sample-path-based Approach. *IEEE/ACM Trans. Netw.* 24, 1 (2016).

Who Falls for Online Political Manipulation?

Adam Badawy
Information Sciences Institute,
University of Southern California
abadawy@usc.edu

Kristina Lerman
Information Sciences Institute,
University of Southern California
lerman@isi.edu

Emilio Ferrara
Information Sciences Institute,
University of Southern California
emiliofe@usc.edu

ABSTRACT

Social media, once hailed as a vehicle for democratization and the promotion of positive social change across the globe, are under attack for becoming a tool of political manipulation and spread of disinformation. A case in point is the alleged use of trolls by Russia to spread malicious content in Western elections. This paper examines the Russian interference campaign in the 2016 US presidential election on Twitter. Our aim is twofold: first, we test whether predicting users who spread trolls' content is feasible in order to gain insight on how to contain their influence in the future; second, we identify features that are most predictive of users who either intentionally or unintentionally play a vital role in spreading this malicious content. We collected a dataset with over 43 million election-related posts shared on Twitter between September 16 and November 9, 2016, by about 5.7 million users. This dataset includes accounts associated with the Russian trolls identified by the US Congress. Proposed models are able to very accurately identify users who spread the trolls' content (average AUC score of 96%, using 10-fold validation). We show that political ideology, bot likelihood scores, and some activity-related account meta data are the most predictive features of whether a user spreads trolls' content or not.

KEYWORDS

Political Manipulation, Russian Trolls, Bots, Social Media

ACM Reference Format:
Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who Falls for Online Political Manipulation?. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316494

1 INTRODUCTION

The initial optimism about the role of social media as a driver of social change has been fading away, following the rise in concerns about the negative consequences of malicious behavior online. Such negative outcomes have been particularly evident in the political domain. The spread of misinformation [29, 33] and the increasing role of bots [4] in the 2016 US presidential elections has increased the interest in automatic detection and prediction of malicious actors.

In this study, we focus on the role of Russian trolls in the recent US presidential election. Trolls are usually described as users who intentionally "annoy" or "bother" others in order to elicit an emotional response. They post inflammatory messages to spread discord and cause emotional reactions [25]. In the context of the 2016 US election, we define trolls as *users who exhibit a clear intent to deceive or create conflict.* Their actions are directed to harm the political process and cause distrust in the political system. Our definition captures the new phenomenon of paid political trolls who are employed by political actors for a specified goal. The most recent and important example of such phenomenon is the Russian "troll farms"—trolls paid by the Russian government to influence conversations about political issues aimed at creating discord and hate among different groups [2, 12].

Survey data from the Pew Research Center [13] show that two-thirds of Americans get their news from Social Media. Moreover, they are being exposed to more political content written by ordinary people than ever before. Bakshy et al. [3] report that 13% of posts by Facebook users—who report their political ideology—are political news. This raises the question of how much influence the Russian trolls had on the national political conversation prior to the 2016 US election, and how much influence such trolls will have in the upcoming elections. Although we do not discuss the effect that these trolls had on the political conservation prior to the election, we focus our efforts in this paper on two questions: **RQ1**: *Can we predict which users will become susceptible to the manipulation campaign by spreading content promoted by Russian trolls?* **RQ2**: *What features distinguish users who spread trolls' messages?*

The goal of these questions is, first, to test whether it is possible to identify the users who will be vulnerable to manipulation and participate in spreading the messages trolls post. We refer to such users as *spreaders* in this paper. Our second goal is to better understand what distinguishes spreaders form non-spreaders. If we can predict who will become a spreader, we can design a counter-campaign, which might stop the manipulation before it achieves its goal.

For this study, we collected Twitter data over a period of seven weeks in the months leading up to the election. We obtained a dataset of over 43 million tweets generated by about 5.7 million distinct users between September 16 and November 9, 2016. First, we cross-referenced the list of Russian trolls published by the US Congress with our dataset and found that 221 Russian trolls exist in our data. Next, we identified the list of users who retweeted the trolls. We gather important features about the users and use a machine learning framework to address the questions posed earlier.

We used different machine learning classifiers on different models (each model includes a subset of the features, with the full model including all the features). We are able to achieve over 90% average AUC score in the full model using Gradient Boosting. In terms of

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316494

feature importance, political ideology is the most prominent feature; number of followers, statuses (no. tweets), and bot scores were also in the top most predictive features. We finally discuss robustness and sanity checks that we carried out to generalize our results.

2 RELATED LITERATURE

The use of trolls and bots in political manipulation campaigns around the globe is well documented through an array of reports by mainstream media outlets and academics (*cf.* Tucker et al. [32] for a comprehensive review). This phenomenon is not entirely new: researchers warned about the potential for online political manipulation for over a decade [15, 17]. Reports tracking and studying this phenomenon date back to the early 2010s [21, 27, 28]. Since then, an increasing account of such events has been recorded in the context of several elections, both in the United States [4, 19, 20, 35, 38, 39] and all over the world, including the U.K. [16], and Italy [6].

Although trolls do not necessarily need to be automated accounts, in many cases bots play a substantial role in exerting influence on social media [22]. Bessi and Ferrara [4] estimated that up to 400K bots were responsible for posting 3.8 million tweets in the last month of the 2016 US presidential election, which is one-fifth of the total volume of online conversations they collected. Russian political manipulation campaigns did not only target the US: there is evidence of Russian interference in German and British elections, and the Catalonian referendum [30]. Russian interference was also reported in the 2017 French presidential election, where bots were detected during the so-called *MacronLeaks* disinformation campaign [9]. Moreover, a recent NATO report claims that around 70% of accounts tweeting in Russian and directed at Baltic countries and Poland are bots.

3 DATA COLLECTION

3.1 Twitter Dataset

We created a list of hashtags and keywords that relate to the 2016 U.S. Presidential election. The list was crafted to contain a roughly equal number of hashtags and keywords associated with each major Presidential candidate: we selected 23 terms, including five terms referring to the Republican Party nominee Donald J. Trump (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump), four terms for Democratic Party nominee Hillary Clinton (#hillaryclinton, #imwithher, #nevertrump, #hillary), and several terms related to debates.

By querying the Twitter Search API continuously and without interruptions between September 15 and November 9, 2016, we collected a large dataset containing 43.7 million unique tweets posted by nearly 5.7 million distinct users. Table 1 reports some aggregate statistics of the dataset. The data collection infrastructure ran inside an Amazon Web Services (AWS) instance to ensure resilience and scalability. We chose to use the Twitter Search API to make sure that we obtained all tweets that contain the search terms of interest posted during the data collection period, rather than a sample of unfiltered tweets. This precaution we took avoids known issues related to collecting sampled data using the Twitter Stream API that had been reported in the literature [23].

Table 1: Descriptive statistics of Twitter data

Statistic	Count
# of Tweets	43,705,293
# of Retweets	31,191,653
# of Distinct Users	5,746,997
# of Tweets/Retweets with a URL	22,647,507

Table 2: Descriptive statistics on Russian trolls

	Value
# of Russian trolls	2,735
# of trolls in our data	221
# of trolls wrote original tweets	85
# of original trolls' tweets	861

Table 3: Descriptive statistics of spreaders

	Value
# of spreaders	40,224
# of times retweeted trolls	83,719
# of spreaders with original tweets	28,274
# of original tweets	>1.5 Million
# of other tweets and retweets	>12 Million

3.2 Russian Trolls

We used a list of 2,752 Twitter accounts identified as Russian trolls that was compiled and released by the U.S. Congress.[1] Table 2 offers some descriptive statistics of the Russian troll accounts. Out of the accounts appearing on the list, 221 exist in our dataset, and 85 of them produced original tweets (861 tweets). Russian trolls in our dataset retweeted 2,354 other distinct users 6,457 times. Trolls retweeted each other only 51 times. Twitter users can choose to report their location in their profile. Most of the self-reported locations of accounts associated with Russian trolls were within the U.S. (however, a few provided Russian locations in their profile), and most of the tweets were from users whose location was self-reported as Tennessee and Texas (49,277 and 26,489 respectively). Russian trolls were retweeted 83,719 times, but most of these retweets were for three troll accounts only: 'TEN_GOP', received 49,286 retweets; 'Pamela_Moore13', 16,532; and 'TheFoundingSon', 8,755. These three accounts make up for over 89% of the times Russian trolls were retweeted. Overall, Russian trolls were retweeted by 40,224 distinct users.

3.3 Spreaders

Users who rebroadcast content produced by Russian trolls, hereafter referred to as *spreaders*, may tell a fascinating story, thus will be the subject of our further investigation. Out of the forty thousand total spreaders, 28,274 of them produced original tweets (the rest only generated retweets). Overall, these 28K spreaders produced over 1.5 Million original tweets and over 12 Million other tweets and retweets—not counting the ones from Russian trolls (cf., Table 3).

4 DATA ANALYSIS & METHODS

In order to answer the questions posed in this paper, we gather a set of features about the users to *(i)* predict the spreaders with the highest accuracy possible and *(ii)* identify feature(s) which best distinguish spreaders from the rest. Table 4 shows all the features we evaluated in this paper, grouped under the following categories:

[1] See Recode.net: https://www.recode.net/2017/11/2/16598312/

Table 4: Features used to describe users in our study

Metadata	LIWC	Engagement	Activity	Other
# of followers	Word Count	Retweet variables	# of characters	Political Ideology
# of favourites	Postive Emotion	Mention variables	# of hashtags	Bot Score
# of friends	Negative Emotion	Reply variables	# of mentions	Tweet Count
Status count	Anxiety	Quote variables	# of urls	
Listed count	Anger			
Default Profile	Sadness			
Geo-enabled	Analytic			
Background-image	Clout			
Verified	Affection			
Account Age	Tone			

Figure 1: Feature correlation for all users in the dataset.

Metadata, Linguistic Inquiry and Word Count (LIWC), Engagement, Activity, and Other variables.

To understand what each variable in the Metadata and LIWC categories means, see the Twitter API documentation and [24], respectively. The Activity variables convey the number of characters, hashtags, mentions, and URLs produced by users, normalized by the number of tweets they post. Tweet Count, under Other, is the number of user's tweets appearing in our dataset. The remaining variables are more involved and warrant a detailed explanation: we explain how Political Ideology, Bot Scores, and Engagement variables were computed in the following sections. One may wonder how much the features evaluated here correlate with each other, and whether they provide informative signals in terms of predictive power about the spreaders. Figure 1 shows that, besides Engagement variables, most of the features are not highly correlated among each other (Pearson correlation is shown, results do not vary significantly for Spearman correlation). There are however a few notable exceptions: Word Count and Tweet Count, LIWC Positive Emotion and Affection, Anxiety and Anger—these pairs all show very high correlation. This is not surprising, considering that these constructs are conceptually close one another. As for the Engagement variables, we can see a "rich get richer" effect here, where users who have higher scores in terms of some of the sub-features in the Engagement category, are also higher in other sub-features. For example, by construction the *Retweet h-index* will be proportional to the number of times a user is retweeted, and similarly for replies, quotes and mentions—all these Engagement features are explained in great detail in a section §4.3.

Table 5: Liberal & Conservative domain names

Liberal	Conservative
www.huffingtonpost.com	www.breitbart.com
thinkprogress.org	www.thegatewaypundit.com
www.politicususa.com	www.lifezette.com
shareblue.com	www.therebel.media
www.dailykos.com	theblacksphere.net

4.1 Political Ideology

4.1.1 Classification of Media Outlets. We classify users by their ideology based on the political leaning of the media outlets they share. We use lists of partisan media outlets compiled by third-party organizations, such as AllSides[2] and Media Bias/Fact Check.[3] The combined list includes 249 liberal outlets and 212 conservative outlets. After cross-referencing with domains obtained in our Twitter dataset, we identified 190 liberal and 167 conservative outlets. We picked five media outlets from each partisan category that appeared most frequently in our Twitter dataset and compiled a list of users who tweeted from these outlets. The list of media outlet domain names for each partisan category is in Table 5.

We used a polarity rule to label Twitter users as liberal or conservative depending on the number of tweets they produced with links to liberal or conservative sources. In other words, if a user had more tweets with links to liberal sources, he/she would be labeled as liberal and vice versa. Although the overwhelming majority of users include links that are either liberal or conservative, we remove any users that had equal number of tweets from each side[4]—this to avoid the conundrum of breaking ties with some arbitrary rule. Our final set of labeled users include 29,832 users.

4.1.2 Label Propagation. We used *label propagation*[5] to classify Twitter accounts as liberal or conservative, similar to prior work [5]. In a network-based label propagation algorithm, each node is assigned a label, which is updated iteratively based on the labels of the node's network neighbors. In label propagation, a node takes the most frequent label of its neighbors as its own new label. The algorithm proceeds updating labels iteratively and stops when the labels no longer change (see [26] for more information). The algorithm takes as parameters *(i)* weights, in-degree or how many times node i retweeted node j; *(ii)* seeds (the list of labeled nodes). We fix the seeds' labels so they do not change in the process, since this seed list also serves as our ground truth.

We construct a retweet network where each node corresponds to a Twitter account and a link exists between pairs of nodes when one of them retweets a message posted by the other. We use the 29K users mentioned in the media outlets sections as seeds, those who mainly retweet messages from either the liberal or the conservative media outlets in Table 5, and label them accordingly. We then run label propagation to label the remaining nodes in the retweet network.

To validate results of the label propagation algorithm, we applied stratified 5-fold cross validation to the set of 29K seeds. We train

[2]See AllSides: https://www.allsides.com/media-bias/media-bias-ratings
[3]See Media Bias Fact Check: https://mediabiasfactcheck.com/
[4]We use five categories, as in left, left center, center, right center, right, to make sure we have a final list of users who are unequivocally liberal or conservative and do not fall in the middle. The media outlet lists for the left/right center and center were compiled from the same sources.
[5]We used the algorithm in the Python implementation of IGraph [7]

Table 6: User breakdown by political ideology

	Liberal	Conservative
# of users	>3.4 M	>1 M
# of trolls	107	108
# of spreaders	1,991	38,233

the algorithm on four-fifths of the seed list and test how it performs on the remaining one-fifth. The averge precision and recall scores are both over 91%.

To further validate the labeling algorithm, we notice that a group of Twitter accounts put media outlet URLs as their personal link/website. We compile a list of the hyper-partisan Twitter users who have the domain names from Table 5 in their profiles and use the same approach explained in the previous paragraph (stratified 5-fold cross-validation). The average precision and recall scores for the two validation methods we use is 91% and 93% respectively, cementing our confidence in the performance of the labeling algorithm.

4.2 Bot Detection

Determining whether either a human or a bot controls a social media account has proven a very challenging task [10, 31]. We use an openly accessible solution called Botometer (a.k.a. BotOrNot) [8, 40], consisting of both a public Web site (https://botometer.iuni.iu.edu/) and a Python API (https://github.com/IUNetSci/botometer-python), which allows for making this determination with high accuracy. Botometer is a machine-learning framework that extracts and analyses a set of over one thousand features, spanning six sub classes:

User: Meta-data features that include the number of friends and followers, the number of tweets produced by the users, profile description and settings.

Friends: Four types of links are considered here: retweeting, mentioning, being retweeted, and being mentioned. For each group separately, botometer extracts features about language use, local time, popularity, etc.

Network: Botometer reconstructs three types of networks: retweet, mention, and hashtag co-occurrence networks. All networks are weighted according to the frequency of interactions or co-occurrences.

Temporal: Features related to user activity, including average rates of tweet production over various time periods and distributions of time intervals between events.

Content: Statistics about length and entropy of tweet text and Part-of-Speech (POS) tagging techniques, which identifies different types of natural language components, or POS tags.

Sentiment: Features such as: arousal, valence and dominance scores [36], happiness score [18], polarization and strength [37], and emotion score [1].

We utilize Botometer to label all the spreaders, and we get bot scores for over 34K out of the total 40K spreaders. Since using Botometer to get scores all non-spreaders (i.e., over 5.7M users) would take an unfeasibly long time (due to Twitter's restrictions), we randomly sample the non-spreader user list and use Botometer to get scores for a roughly equivalent-size list of non-spreader users. The randomly-selected non-spreader list includes circa 37K users. To label accounts as bots, we use the fifty-percent threshold which has proven effective in prior studies [8]: an account is considered to be a bot if the overall Botometer score is above 0.5. Figure 2 shows the probability distribution for spreaders vs. non-spreaders. While most of the density is under the 0.5 threshold, the mean of spreaders (0.3) is higher than the mean of non-spreaders. Additionally, we used a t-test to verify that the difference is significant at the 0.001 level (p-value).

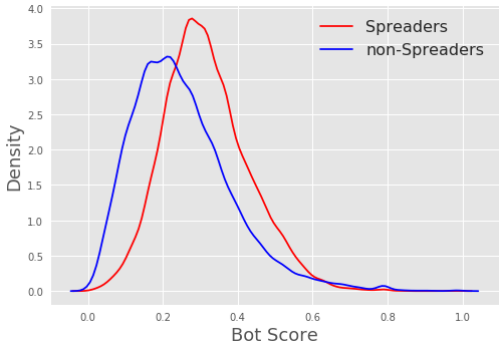


Figure 2: Probability density distributions of bot scores assigned to spreaders (red) and non-spreaders (blue).

As for the plots in Figure 3, it is evident that the spreaders are different on almost all the Botometer subclass scores, except for the temporal features. The differences in all plots are statistically significant ($p < 0.001$). Besides, looking at the distributions, we can see that the difference in user characteristics (metadata), friends, and network distributions, are substantively different as well. Moreover, the mean of spreaders is higher in all the subclass features.

4.3 Engagement

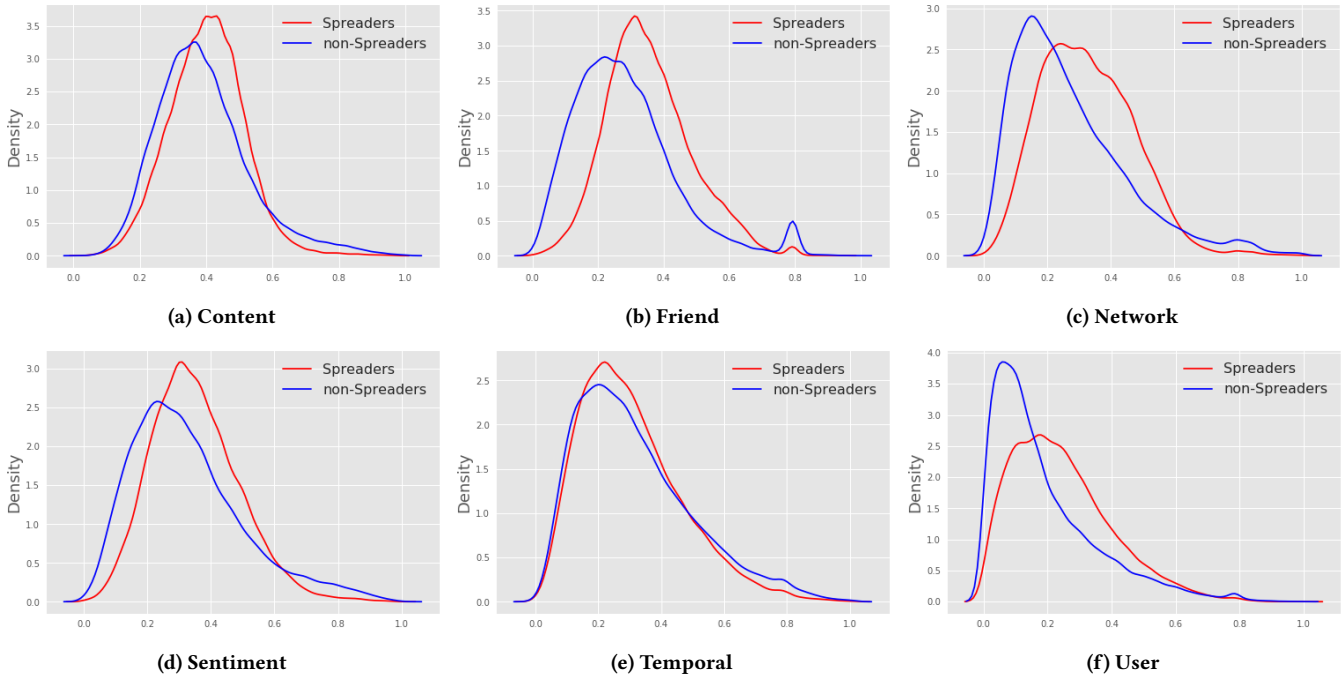
We plan to measure user engagement in four activities: retweets, mentions, replies, and quotes. Engagement of a user is measured through three components: the quantity, longevity, and stability in each activity. For instance, for a set of N users, this measure would calculate the engagement index score of user $i \in N$ by including the following:

1) number of retweets, replies, mentions, and quotes by $N - i$ users for user i;
2) time difference between the last and the first quote, reply, and retweet per tweet;
3) consistency of mentioning, replying, retweeting, and quoting by $N - i$ users for user i across time (per day);
4) number of unique users who retweeted, commented, mentioned, and quoted user i

Item three is measured using h-index [14]. The measure captures two notions: how highly referenced and how continuously highly referenced a user is by other members in the network. This measure was originally proposed to quantify an individual's scientific research output. In this context, a user has index h if for h days, he/she is referenced at least h times and in all but h days no more than h times.

5 RESULTS

Predicting spreaders on the original dataset may be considered a daunting task: only a relatively small fraction of users engaged



(a) Content **(b) Friend** **(c) Network**

(d) Sentiment **(e) Temporal** **(f) User**

Figure 3: Distribution of the probability density of Botometer subfeature scores for spreaders vs. nonspreaders.

with Russian trolls' content (about 40K out of 5.7M users). However, for the same reason, if a model were to trivially predict that no user will ever engage with Russian trolls, the model would be accurate most of the time (i.e., most users won't be spreaders), even if its recall would be zero (i.e., the model would never correctly predict any actual spreaders)—provided that we want to predict spreaders, this model would not be very useful in practice. In other words, our setting is a typical machine-learning example of a highly-unbalanced prediction task.

To initially simplify our prediction task, we created a balanced dataset that is limited to users who have bot scores. Will get back to the original prediction task on the highly-unbalanced dataset later. This balanced dataset has about 72K users, with 34K spreaders and 38K non-spreaders. To test our ability to detect spreaders and to see which features are most important in distinguishing between the two groups, we leverage multiple classifiers and multiple models: the first model serves as a baseline with each model including more variables until we reach the full model. Since our goal was not that to devise new techniques, we used four off-the-shelf machine learning algorithms: Extra Trees, Random Forest, Adaptive Boosting, and Gradient Boosting. We train our classifiers using Stratified 10-fold cross-validation with the following preprocessing steps *(i)* replace all categorical missing values with the most frequent value in the column *(ii)* replace missing values with the mean of the column.

Table 7 shows all the models we evaluate, from the simplest baseline model (Metadata) to the full model that includes all the features we present in Table 4.

For Gradient Boosting, which is the best performing classifier among the four we evaluate, we obtained average AUC scores for the 10 folds that range from 85% to 96%. Figure 4 shows the ROC curve plots for each model (using the fold/model with the highest

Table 7: Models: from Baseline (Metadata) to Full

Model	Features
1	Metadata
2	Metadata + LIWC
3	Metadata + LIWC + Activity
4	Metadata + LIWC +Activity + Engagement
5	Metadata + LIWC +Activity + Engagement + Other

AUC score among the trained ones). The jump from 89% to 96% for the AUC scores from Model 4 to 5 shows that the addition of bot scores and political ideology are meaningful in distinguishing spreaders from non-spreaders (the legend in Figure 4 shows the average AUC score for each model). To better understand the contribution of the features in predicting the target values (i.e., spreader vs. non-spreader), we look at the variable importance plot of the Gradient Boosting results for Model 5. The *Variable Importance* plot (cf., Figure 5) provides a list of the most significant variables in descending order by a mean decrease in the Gini criterion. The top variables contribute more to the model than the bottom ones and can discriminate better between spreaders and non-spreaders. In other words, features are ranked based on their predictive power according to the given model. Figure 5 shows that, according to Model 5 and Gradient Boosting, political ideology is the most predictive feature, followed by number of followers, statuses/tweets count (obtained from the metadata), and bot score, in a descending order of importance. The plot does not show all the features, since the omitted features contribute very little to the overall predictive power of the model.

Feature importance plots reveal which features contribute most to classification performance, but they do not tell us the nature of the relationship between the outcome variable and the predictors. Although predictive models are sometime used as black boxes,

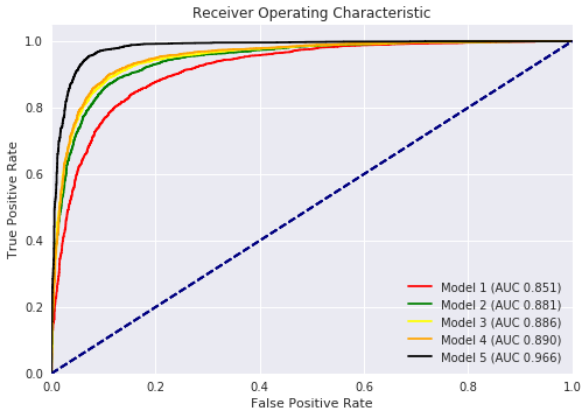


Figure 4: Area under the ROC curve plot for five GB models

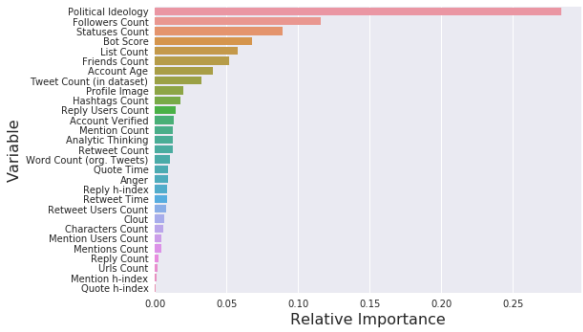
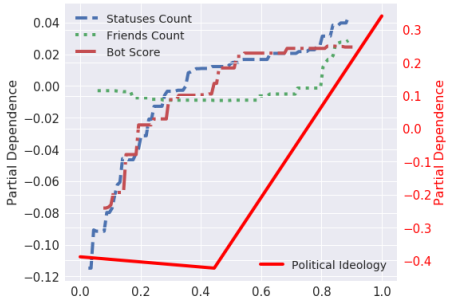


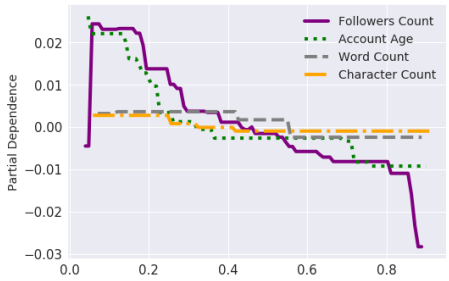
Figure 5: Relative importance of the features (GB full model)

Partial Dependence plots (cf., Fig. 6) can tell us a lot about the structure and direction of the relationship between the target and independent variables. They show these relationships after the model is fitted, while marginalizing over the values of all other features. The dependency along the x-axis captures the range of each feature, with that feature values normalized between 0 and 1. Political ideology should be considered in the range from 0 (to identify left-leaning users), to 1 (for right-leaning ones).

Using Partial Dependence, we illustrate that the target variable (spreader) has positive relationships with the following features: political ideology, statuses count, bot scores, and friends count. Figure 6a visualizes these relationships (we put political ideology on a different y-axis in order to show that its magnitude of influence on the target variable is significantly higher compared to all other features, including downward trend features in Figure 6b). This suggests that moving from left to right political leaning increases the probability of being a spreader; larger number of posts, more friends, and higher bot scores are also associated with higher likelihood of being a spreader. On the other hand, we can see that the outcome variable has a negative relationship with followers count, account age, characters count, and word count, as shown in Figure 6b. This means that having fewer followers, having a recently-created account, posting shorter tweets with fewer words, are all characteristics associated with higher probability of being a spreader.



(a) Upward Trends



(b) Downward Trends

Figure 6: Partial Dependence plots for some important features (GB full model). Plots show the dependence of the outcome variable (spreader) on each feature, marginalizing over all the other features (x-axis values are CDF-normalized).

5.1 Robustness and Generalization Analysis

Original vs. Rebalanced Data: Going back to the original highly-unbalanced dataset, we aim to validate the results above using two strategies: *(i)* we run Gradient Boosting (with the same preprocessing steps) on the whole dataset of 5.7M for the five models we outlined in table 7; *(ii)* we run Gradient Boosting on models without imputations and with all missing observations deleted. For the first approach, the average AUC scores ranged form 83% for the baseline model to 98% for the full model. In terms of feature importance, bot score and political ideology are the most predictive features as expected in the full model prediction. For the second approach, due to the sparsity of some features, the overall number of observations decreases significantly when these features are added. Putting the overall number of observations aside, the average ROC scores for a 10-fold validation for the roughly same set of models specified earlier range from 84% to 91%. In terms of feature importance, political ideology is the most important feature in the full model, with status count and bot scores following. In summary, the results above remain consistent when validating on the highly-unbalanced prediction.

Excluding bots: Removing Bots from the balanced dataset then predicting spreaders using Gradient Boosting yields the same AUC scores for models 1-5, ranging from 85% to 96%. In terms of feature selection, political ideology is the best predictive feature by a wide margin then list, followers, posts (tweets), and friends count in this exact order. Although the ordering is different for the balanced dataset with bots, these features are the top 4 (if we exclude bot scores) most predictive features with or without the inclusion of bots.

Error Analysis: Few trends emerge when we compare correctly labeled users vs the ones that are misclassified in the balanced dataset. Correctly labeled users, in terms of metadata, have more followers, favourites, friends, and posts. Moreover, they have higher variance in terms of word count and anger words and they use more analytical words.

6 DISCUSSION AND CONCLUSION

The results in previous section show that *(i)* with some insight on users who spread or produce malicious content, we are able to predict those that will spread trolls' messages; *(ii)* in the case of the 2016 US presidential election, political ideology was highly predictive of who is going to spread trolls' messages. Fig. 3b, 3c, and 3f show that spreaders and non-spreaders are significantly different on the dimensions of friends, network, and user metadata, with spreaders having higher bot scores on all three. Basic metadata features give a strong signal in terms of differentiating spreaders from non-spreaders, along with the bot score (cf. Fig. 5).

Looking at the partial dependence plots, we can deduce that spreaders write a lot of tweets (counting retweets as well), have higher bot scores, and tend to be more conservative (conservative is labeled as the highest numerical value in the political ideology feature). Also, since the range of the y-axis tells us about the range of influence a feature has on the target value, it is evident that political ideology has by far the most influence on distinguishing between spreaders and non-spreaders. On the other hand, we can also deduce that spreaders do not write much original content, tend not have that many followers, and have more recently established user accounts. In the downward trends in Figure 6b, we can see that followers count and account age have more influence on the target value in comparison to the other features in this plot. To conclude, this paper focused on predicting spreaders who fall for online manipulation campaigns. We believe that identifying likely victims of political manipulation campaigns is the first step in containing the spread of sponsored content [11, 34]. Access to reliable and trustworthy information is a cornerstone of any democratic society. Declining trust of citizens of democratic societies in mainstream news and their increased exposure to content produced by ill-intended sources poses a danger to democracy.

Acknowledgements. This work was supported by the Air Force Office of Scientific Research (#FA9550-17-1-0327).

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, et al. 2011. Sentiment analysis of twitter data. In *ACL*. 30–38.

[2] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *ASONAM*. 258–265.

[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* (2015).

[4] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 11 (2016).

[5] M Conover, B Gonçalves, J Ratkiewicz, et al. 2011. Predicting the Political Alignment of Twitter Users. In *Social Computing*. 192–199.

[6] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, et al. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *WWW*. 963–972.

[7] Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* (2006).

[8] Clayton Allen Davis, Onur Varol, Emilio Ferrara, et al. 2016. Botornot: A system to evaluate social bots. In *WWW*. 273–274.

[9] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22 (2017).

[10] Emilio Ferrara, Onur Varol, Clayton Davis, et al. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[11] Emilio Ferrara, Onur Varol, Filippo Menczer, et al. 2016. Detection of promoted social media campaigns. In *ICWSM*. 563–566.

[12] Theodore P Gerber and Jane Zavisca. 2016. Does Russian propaganda work? *The Washington Quarterly* 39, 2 (2016), 79–98.

[13] Jeffrey Gottfried and Elisa Shearer. 2016. *News Use Across Social Medial Platforms 2016*. Pew Research Center.

[14] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *PNAS* 102, 46 (2005), 16569.

[15] Philip Howard. 2006. *New media campaigns and the managed citizen*.

[16] Philip Howard and Bence Kollanyi. 2016. Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. (2016).

[17] Tim Hwang, Ian Pearce, and Max Nanis. 2012. Socialbots: Voices from the fronts. *Interactions* 19, 2 (2012), 38–45.

[18] Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, et al. 2012. Positivity of the English language. *PloS one* 7, 1 (2012).

[19] Bence Kollanyi, Philip Howard, and Samuel Woolley. 2016. Bots and automation over Twitter during the first US Presidential debate. (2016).

[20] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *Data & Society Research Institute* (2017).

[21] Panagiotis T Metaxas and Eni Mustafaraj. 2012. Social media and the elections. *Science* 338, 6106 (2012), 472–473.

[22] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, et al. 2017. Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots. *Plos One* 12, 9 (2017), e0184148.

[23] Fred Morstatter, Jürgen Pfeffer, Huan Liu, et al. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*. 400–408.

[24] James Pennebaker, Ryan Boyd, Kayla Jordan, et al. 2015. The development and psychometric properties of LIWC2015. (2015).

[25] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*.

[26] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.

[27] Jacob Ratkiewicz, Michael Conover, Mark Meiss, et al. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *WWW*.

[28] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, et al. 2011. Detecting and tracking political abuse in social media. *ICWSM* (2011).

[29] Samantha Shorey and Philip Howard. 2016. Automation, Algorithms, and Politics: A Research Review. *Int. J Comm.* 10 (2016).

[30] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* (2018).

[31] VS Subrahmanian, Amos Azaria, Skylar Durst, et al. 2016. The DARPA Twitter bot challenge. *Computer* 49, 6 (2016).

[32] Joshua Tucker, Andrew Guess, Pablo Barberá, et al. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. (2018).

[33] Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, et al. 2017. From liberation to turmoil: social media and democracy. *Journal of democracy* 28, 4 (2017), 46–59.

[34] Onur Varol, Emilio Ferrara, Filippo Menczer, et al. 2017. Early Detection of Promoted Campaigns on Social Media. *EPJ Data Science* 6, 13 (2017).

[35] Yu Wang, Yuncheng Li, and Jiebo Luo. 2016. Deciphering the 2016 US Presidential Campaign in the Twitter Sphere: A Comparison of the Trumpists and Clintonists. In *ICWSM*. 723–726.

[36] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

[37] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP*.

[38] Samuel C Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday* 21, 4 (2016).

[39] Samuel C Woolley and Philip Howard. 2016. Automation, Algorithms, and Politics: Introduction. *Int. Journal of Commun.* 10 (2016).

[40] Kai-Cheng Yang, Onur Varol, Clayton A Davis, et al. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* (2019), e115.

Robust Cyberbullying Detection with Causal Interpretation

Lu Cheng
Computer Science and Engineering,
Arizona State University
lcheng35@asu.edu

Ruocheng Guo
Computer Science and Engineering,
Arizona State University
rguo12@asu.edu@asu.edu

Huan Liu
Computer Science and Engineering,
Arizona State University
huanliu@asu.edu

ABSTRACT

Cyberbullying poses serious threats to preteens and teenagers, therefore, understanding the incentives behind cyberbullying is critical to prevent its happening and mitigate the impact. Most existing work towards cyberbullying detection has focused on the accuracy, and overlooked causes of the outcome. Discovering the causes of cyberbullying from observational data is challenging due to the existence of *confounders*, variables that can lead to spurious causal relationships between covariates and the outcome. This work studies the problem of robust cyberbullying detection with causal interpretation and proposes a principled framework to identify and block the influence of the plausible confounders, i.e., *p*-confounders. The de-confounded model is causally interpretable and is more robust to the changes in data distribution. We test our approach using the state-of-the-art evaluation method, *causal transportability*. The experimental results corroborate the effectiveness of our proposed algorithm. The purpose of this study is to provide a computational means to understanding cyberbullying behavior from observational data. This improves our ability to predict and to facilitate effective strategies or policies to proactively mitigate the impact of cyberbullying.

KEYWORDS

Cyberbullying detection, Causality, Social media

ACM Reference Format:
Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust Cyberbullying Detection with Causal Interpretation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316503

1 INTRODUCTION

Electronic communication has provided a new context for preteen and teenage youth to bully and harass their peers. The prevention of cyberbullying behavior, hence, has been a growing public concern in the US and abroad [9]. However, preventing cyberbullying and its related harms is complicated with the limited understanding of the causes. Within the data mining community, primary efforts toward automatic detection of cyberbullying have focused on building generic binary classifiers with high accuracy rather than causal interpretation. As data for cyberbullying detection on various social media platforms are hard to obtain, this requires that cyberbullying classifiers are robust to *confounding bias* and *selection bias* to ensure the validity of studies based on the binary classification.

Even though confounder is central to much of empirical social science, it has been mostly overlooked in standard machine learning models. This is presumably because *prediction* is the goal of machine learning models, instead of *causation*. Despite their satisfactory performance, these models fail to explain the *causes* behind each prediction. The strong statistical association between the input covariates (e.g., text features) and the outcome (class label) can benefit from another covariate that is correlated with both the input and the output. For example, previous studies [17] reveal the strong correlation between Facebook intensity (a measurement of Facebook usage) and cyberbullying, but does this imply that high Facebook intensity cause cyberbullying? Studies have shown that this is spurious association due to the influence of user's age and gender, i.e., confounders [17]. Understanding confounders can also encourage generalizable science [21]. For cyberbullying detection, as data is hard to collect, conclusions that are obtained in one experimental setting should be effectively applied to another similar environment, i.e., *causal transportability* [21]. Indeed, if influence of confounders is consistent from training to testing data, the predictions should not be harmed by the presence of confounders. However, this is not true in practice as on one hand, training sets are typically small due to the annotation cost, which tends to change the data distributions. On the other hand, in many domains, relationship between the confounder and the outcome is likely to shift over time, resulting in poor accuracy. Without proper control of confounding bias, we can easily reach problematic conclusions.

Building a robust cyberbullying classifier remains a challenging task, mainly because: (1) It is often impossible to identify causal structures from pure observational data, where considerable uncertainty presents in the data generating process. A common relaxation is to find the covariates that are associated with the outcome conditioning on other observed covariates [12]. (2) Online data from social media platforms is often sparse, noisy, heterogeneous and high-dimensional, leading to tremendous spurious associations among covariates. While researchers have analyzed growing social data to understand social behavior in online platforms, it presents multi-faceted challenges that current machine learning models are not well-equipped to handle [1]. To achieve the goal of robust cyberbullying detection with causal interpretation, in this paper, we study a novel problem of controlling confounding bias in cyberbullying detection based on observed social media data. This essentially enables us to identify *plausible* confounders (*p*-confounder) and covariates that are causally related to cyberbullying behavior.

The contributions of this work are highlighted below:
(1) We study the problem of robust cyberbullying detection via the identification and control of *p*-confounders. To the best of our

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316503

knowledge, this is the first work aimed at leveraging the connections between machine learning and causality to build a robust and causally interpretable cyberbullying classifier.
(2) We develop a principled statistical method to identify p-confounders in cyberbullying detection. To block influence of these confounders, we apply clustering algorithm to stratify data into more homogeneous subgroups and classify new instances accordingly.
(3) We perform extensive experiments on two real-world datasets to examine models' robustness based on *causal transportability*, the state-of-the-art evaluation method of learning confounders from the observational data.

2 PROBLEM STATEMENT

Given a corpus of N social media posts, we denote as $\mathbf{y} = \{y_1, ..., y_N\}$ the labels of the posts with $y_i \in \{0, 1\}, i = \{1, 2, ..., N\}$, where $y_i = 1$ denotes that the post is a cyberbullying message, otherwise $y_i = 0$. $\mathbf{X}$ is the covariate matrix and its i-th column, X^i denotes the i-th covariate. The problem of robust cyberbullying detection with causal interpretation is to identify and block the influence of a group of p-confounders $\mathcal{M} = \{..., X^m, ...\}, X^m \in \mathcal{X}$ in cyberbullying detection.

To this end, we aim to answer:
How to effectively detect and block the influence of plausible confounders from the observational cyberbullying data?
Following the convention in the literature [29], we first address two required assumptions:

- *Unconfoundedness assumption.* All the covariates affecting both the input covariates X and the cyberbullying outcome Y are observed.
- *No complex causes.* A cause of cyberbullying behavior is as simple as being represented as a single covariate.

These assumptions are important as they are the foundations of the general causal inference framework. It explains that we do not consider uncertainty in the observational data due to unobserved confounders. It also assumes that no cause is composed of other causes.

3 PROPOSED FRAMEWORK

In this section, we describe the details of how to find confounder candidates and identify p-confounders.

3.1 Generate confounder candidates

Cyberbullying study can be susceptible to confounding bias, where a statistical association is distorted and does not reflect the corresponding causal relation. Confounding bias can lead to Simpson's paradox, a phenomenon whereby the association coefficient between a covariate and the outcome reverses sign upon conditioning on a second covariate, regardless of the value of the second covariate [20]. Simpson's paradox is often presented as a compelling demonstration of the existence of confounders [19] as it implies that there might be multiple causal pathways leading to the outcome. For cyberbullying detection, one of the two possible causal graphs (we assume there exist causes among the input covariates) that can result in Simpson's paradox is shown in Fig. 1. Here M is a confounder because it is on the *back-door path* $X \leftarrow M \rightarrow Y$. Hence, the causal relation between X and Y is spurious as M can influence both the main covariate X and the outcome Y through $P(X|M)$ and $P(Y|M)$.

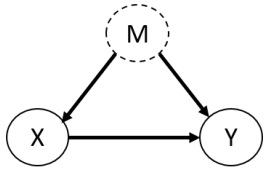


Figure 1: A causal graph leads to the Simpson's paradox where M is the confounder of the causal relationship $X \rightarrow Y$. X and Y are the main covariate and the outcome, respectively.

To detect Simpson's paradox, we first compute the statistical associations between $X^s(X)$ and Y with linear generalized linear model:

$$\mathbb{E}[Y|X^s] = f(\gamma + \beta_1 x^s), \tag{1}$$

where β_1 measures the effect of the main covariate X^s on Y, and f is a monotonically increasing function (e.g. logistic sigmoid function in this work) of its argument. Secondly, we seek for Simpson's pairs $(X^m(M), X^s)$ such that the association between X^s and Y reverses sign upon conditioning on X^m, i.e., Simpson's paradox. Specifically, conditioning on the second covariate X^m, we stratify the data to investigate the trend in each subgroup $g \in G$:

$$\mathbb{E}_c[Y|X^s, X^m] = f(\gamma + \beta_{cg} x^s). \tag{2}$$

Then the following equation should hold to ensure the reversal:

$$\begin{aligned} &\frac{d}{dX^s}\mathbb{E}[Y|X^s]\times \\ &\frac{d}{dX^s}\mathbb{E}_c[Y|X^s, X^m = x^m] < 0 \qquad \forall x^m, \end{aligned} \tag{3}$$

where the first term stands for the gradients of models at the aggregate level of data and the second term is the gradient of model in subgroups conditioning on X^m. Next, to quantify how significant these variables are correlated to each other, we perform a goodness of fit test at the significance level $\alpha = 0.05$. The test statistic of the logistic regression using the Wald test [33] is simply

$$Z = \frac{\hat{\beta}_i}{s.d.(\hat{\beta}_i)} \sim \mathcal{N}(0, 1) \quad \beta_i \in \{\beta_1, \beta_{cg}\}, \tag{4}$$

where $s.d.$ is the standard deviation. Then the obtained p-value determines "likel" or "unlikely" to reject the null hypothesis $\beta_i = 0$, i.e., if the p-value is less than (or equal to) α, then β_i is statistically different from zero and vice versa.

In summation, to observe Simpson's paradox, we need to ensure that

- Reversed sign: $\beta_1 \times \beta_{cg} < 0, \forall g \in G$, satisfying Eq. (3).
- Goodness-of-fit test: $p_{\beta_1} < \alpha, p_{\beta_{cg}} < \alpha, \forall g \in G$, where p_β is the p-value of β.

Now $\mathcal{P}$ includes all identified Simpson's pairs (X^m, X^s) in cyberbullying detection and any X^m appears in a Simpson's pair is a confounder candidate.

3.2 Detect p-confounders

As explained in [19], Simpson's paradox can result from multiple causal graphs, i.e., the presence of Simpson's paradox is a necessary

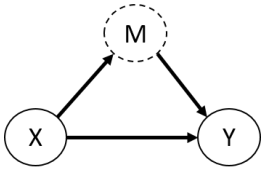


Figure 2: The other causal graph that results in Simpson's paradox in cyberbullying detection. M is not a confounder.

but not sufficient condition for the existence of confounders. In the problem setting of cyberbullying detection, two causal graphs, as shown in Fig. 1-2, can lead to Simpson's paradox as the causal directions between M and Y, X and Y can only be $X \rightarrow Y$ and $M \rightarrow Y$. Comparing the two causal graphs, M in Fig. 1 is the confounder as it influences both X and Y while M in Fig. 2 is not.

This brings up the critical question for us to find out the more likely confounders in cyberbullying detection: *Which causal graph does a Simpson's paradox may imply?* In other words, should we block influence of any variable that can lead to Simpson's paradox? The answer is "No" because there are no back-door paths requiring blockage in Fig. 2.

Let z_{mp} indicate whether the covariate X^m is a confounder candidate, that is

$$z_{mp} = \begin{cases} 1, & \text{if } X_m \text{ is in Simpson's pair } p, \\ 0, & \text{otherwise.} \end{cases}$$

We first define the *confounder momentum* Z^m of a covariate X^m as

$$Z^m = \sum_{p=1}^{|\mathcal{P}|} z_{mp}, \qquad \forall X^m \in \mathcal{X}. \tag{5}$$

We then rank the potential confounders based on the *confounder momentum* Z^m in a descending order. Then the covariates with higher confounder momentum are more plausible confounders. This is intuitive as confounders are defined by the underlying causal structures. A confounder in a causal graph may not be a confounder in another causal graph as input covariate changes from X to X'. When blocking the influence of p-confounders, we seek for confounder candidates that have impact on larger group of input covariates in the observed data.

To identify the more likely causal graph behind a Simpson's paradox, we further propose a data-driven approach based on the theory of *causal transportability* [21]. The crucial question to address here is can we "transport"- or generalize- the cyberbullying classifiers from one population to the other? As predictions relying on causal relationships (de-confounded models) should be more robust to changes of data distribution, we compare the performance of regular cyberbullying classifiers and de-confounded cyberbullying classifiers through across-domain predictions. In the rest of this section, we describe the details of de-confounding procedure and how causal interpretation can be naturally revealed.

Table 1: Dataset Statistic

Dataset	#Users	#Normal	#Bully	#Total
Formspring	50	12,036	1,126	13,162
Twitter	9,833	16,149	3,845	19,994

3.3 Identify potential causes

Conventional machine learning models seek features that can predict well. Nevertheless, this can constrain models' ability of generalization. An intuitive way to block the spurious statistical associations (de-confounding) is data stratification, i.e., disaggregating data into subgroups where more homogeneous instances are grouped together. Given a set of p-confounders $\mathcal{M}$, the data stratification procedure is

$$h : \mathcal{M} \rightarrow g, \ g \in \mathcal{G}, \tag{6}$$

where $\mathcal{G}$ is a set of clusters and h is a clustering function (e.g., Gaussian Mixture Clustering) that assigns samples with similar values of p-confounders to the same cluster. This step is to group instances with more homogeneous values of p-confounders to block their influence on both the covariates and the outcome. In the testing phase, given a new instance, it is first assigned to the closest cluster and then we use the classifier in that subgroup to predict the label for this instance. The de-confounded cyberbullying classifier is causally interpretable as it relies on causal relationships instead of statistical associations. At last, we obtain covariates with large coefficients. The top covariates of the de-confounded cyberbullying classifiers play central roles in causally interpreting the predicted results.

4 DATASETS

We consider two real-world datasets for empirical evaluation of the proposed method: the Formspring[1] dataset and the Twitter dataset. Formspring, as well as Twitter, has been rated as the top social media tool where cyberbullying most frequently occurs[2]. We crawled the Twitter dataset via the Twitter streaming API [15] from September 19th to 25th (2017) with the following keywords as suggested by [16]: *nerd, gay, loser, freak, emo, whale, pig, fat, wannabe, poser, whore, should, die, slept, caught, suck, slut, live, afraid, fight, pussy, cunt, kill, dick, bitch.* 20,000 tweets were manually labeled by two psychologists and a third expert was asked to resolve the conflicts. The initial agreement of the two annotators is 80%. After conflict resolution and data cleaning, we finally obtained the Twitter dataset with 19,994 labeled tweets. Table 1 shows the basic statistics of the two datasets. Code and dataset will be released upon the acceptance of the manuscript.

For the input covariates, we adopt the representation of Linguistic Inquiry and Word Count (LIWC) [22] developed for psychometric analysis. Specifically, LIWC counts words that belong to certain categories in psychology. For example, the word "cry" belongs to five categories: sadness, negative emotion, overall affect, verb and past tense verb. The results of previous research show that such

[1] Available at http://www.chatcoder.com/DataDownload

[2] http://www.foxnews.com/tech/2010/10/07/meanest-sites-prevent-cyberbullying-online-kids.html

Table 2: A sample of LIWC output variable information.

Category	Abbrev	Examples	#words
Affective processes	affect	happy,cried	1,393
Positive emotion	posemo	love,nice	620
Cognitive processes	cogproc	cause,know	797

psychometric analysis can improve the performance of cyberbullying detection [16]. For each data point, the output of LIWC is a numerical vector with covariates from 100 subcategories such as psychological processes and cognitive processes. Table 2 shows a sample of variable information output from LIWC[3].

5 RESULTS

In this section, we design experiments to compare the causal transportability of regular and de-confounded cyberbullying classifiers, and also present some qualitative analyses for further illustration. In particular, the model is trained on one dataset and tested on the other. The experiments are executed on Python 2.7 and R 3.5.1 using a Mac OS X system with Intel Core i5 and 8GB of RAM. The following experimental results are reported with 10 runs on the two datasets.

5.1 Evaluation of p-confounders

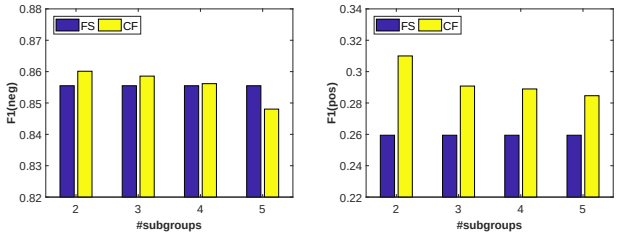
The motivation to detect p-confounders is to remove the spurious associations and identify the underlying causal mechanism of cyberbullying detection. A de-confounded model is also more robust to confounding bias and can be generalized to real-world applications [18]. As it typically takes lots of efforts to obtain real-world data for studying cyberbullying, identification of covariates that are causally related to cyberbullying can facilitate related studies on different social media platforms and digital communication tools.

To evaluate the effectiveness of the p-confounders (covariates with high confounder momentum), we compare the causal transportability of the model at the aggregate level to that at the disaggregate level of data. Specifically, we first use the K-means clustering algorithm to separate the data into K different subgroups based on the p-confounders. Then for a given testing instance, the first step is to find its most similar subgroup and then to make prediction using the classifier trained in that subgroup. The training dataset and the testing dataset should come from different domains but are sufficiently similar to each other, e.g., Twitter and Formspring.

But how to set K, the number of subgroups? The question is critical to answer as it decides how the p-confounders will be controlled and how new data instances will be classified. Here, we investigate the influence of K on the causal transportability of the de-confounded classifiers and decide the optimal K. We employ Random Forest algorithm and set K to {2,3,4,5}. We compare the results to the regular Random Forest classifier (without de-confounding) and report F1 scores for both positive (bullying) cases and negative (normal) cases due to the data imbalance. Fig. 3-4 show the results of across domain cyberbullying detection.

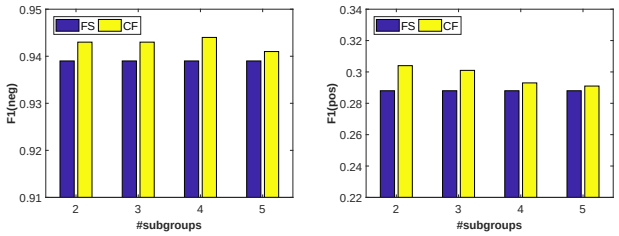
The results show that classifier with de-confounding (CF) outperform the regular classifier (FS) in most cases even when CF is

[3]Details can be seen at http://liwc.wpengine.com/



(a) F1 score (neg) **(b) F1 score (pos)**

Figure 3: Analysis of #subgroups: Formspring → Twitter



(a) F1 score (neg) **(b) F1 score (pos)**

Figure 4: Analysis of #subgroups: Twitter → Formspring

trained on less data due to data disaggregation. The advantages are gradually narrowed down as K increases as data disaggregation sacrifices the number of data for training. The trade-off between the number of training instances and confounding bias control should be studied in future work. Another strength of the de-confounded Random Forest classifier is it shows significant power in the prediction of bullying cases even though the training dataset is imbalanced with much more negative instances. This indicates that data stratification conditioning on p-confounders effectively mitigate influence of confounders and thus the de-confounded model is more robust to the domain changes.

With K set to 2, another research question to be addressed is *how many p-confounders are there?* Here we propose a data-driven approach. Given a ranked list of potential confounders, we use top t confounder candidates to disaggregate data. We then set $t = \{1, 2, 4, 8\}$ and test the causal transportability with three common classification models: Extra Trees classifier that fits a number of randomized decision trees (a.k.a. extra-trees) [14], Random Forest [3] and AdaBoosting [25]. Results are shown in Table 3-4, where CF denotes as the de-confounded classifiers. We can observe from the results that

- In most of the cases, the causal transportability of classifiers trained on each strata of the disaggregated data outperform regular machine learning models. For example, the de-confounded Random Forest algorithm in Table 3 outperforms the regular Random Forest by up to 23.6% in predicting bullying cases. This reveals that (1) covariates with high confounder momentum are p-confounders and (2) the proposed algorithm can effectively block influence of detected p-confounders as well.

Table 3: Robustness comparisons w.r.t. #confounders: Formspring → Twitter

Classifiers		Random Forest				Extra Tree				AdaBoost			
#confounders		1	2	4	8	1	2	4	8	1	2	4	8
F1 (neg)	FS	0.856	0.856	0.856	0.856	0.860	0.860	0.860	0.860	**0.858**	**0.858**	**0.858**	**0.858**
	CF	**0.858**	**0.862**	**0.864**	**0.857**	**0.862**	**0.861**	**0.862**	**0.861**	0.847	0.852	0.846	0.854
F1 (pos)	FS	0.271	0.271	0.271	0.271	0.262	0.262	0.262	0.262	0.334	0.334	0.334	0.334
	CF	**0.280**	**0.311**	**0.335**	**0.315**	**0.270**	**0.286**	**0.297**	**0.297**	**0.359**	**0.371**	**0.372**	**0.346**

Table 4: Robustness comparisons w.r.t. #confounders: Twitter → Formspring

Classifiers		Random Forest				Extra Tree				AdaBoost			
#confounders		1	2	4	8	1	2	4	8	1	2	4	8
F1 (neg)	FS	0.942	0.942	0.942	0.942	0.938	0.938	0.938	0.938	0.918	0.918	0.918	0.918
	CF	**0.943**	**0.945**	**0.945**	**0.944**	**0.944**	**0.942**	**0.941**	**0.942**	**0.934**	**0.935**	**0.934**	**0.930**
F1 (pos)	FS	0.298	0.298	0.298	0.298	0.275	0.275	0.275	0.275	0.311	0.311	0.311	0.311
	CF	**0.308**	**0.309**	**0.309**	**0.306**	**0.291**	**0.292**	**0.326**	**0.299**	**0.326**	**0.328**	**0.333**	**0.332**

Table 5: Robustness comparisons: Covariates with highest confounder momentum VS. lowest ones

#confounders		1	2	4	8
Formspring→Twitter	CF	**0.270**	**0.286**	**0.297**	**0.297**
	LC	0.265	0.265	0.267	0.262
Twitter→Formspring	CF	**0.291**	**0.292**	**0.326**	**0.299**
	LC	0.279	0.277	0.285	0.280

Table 6: Top five important covariates of models in Group VS. Subgroups: LIWC categories Informal, Drives, Affective processes and Biological processes .

Group	Subgroup1	Subgroup2	Subgroup3
swear	power	sexual	anx
anger	achieve	bio	drives
informal	drives	negemo	affect
negemo	health	swear	affiliation

- While the gain from de-confounding on predicting the negative class is not obvious, the performance of cross-domain cyberbullying detection for the positive class is improved significantly. The cyberbullying datasets in our experiments, as well as in the real world, are typically imbalanced. With the number of negative (normal) instances much larger than that of the positive (bully) instances, it is more challenging to train a classifier with high true positive than that with high true negative. However, our experiments indicate that by controlling the confounding bias, the model can find the covariates that are causally related to cyberbullying behavior. Hence it boosts the model's performance of effectively identifying cyberbullying cases.
- As the number of p-confounders used for data disaggregation increases, the de-confounded model is more capable of predicting positive cyberbullying instances. The optimal t is between 4 and 8, hence the number of p-confounders is set to $t = [4, 8)$. We conclude that covariates with large confounder momentum are effective p-confounders and blocking the influence of these covariates can improve the performance of cyberbullying detection.

To further test the effectiveness of our algorithm, we randomly pick {1,2,4,8} covariates with lowest confounder momentum (LC) and repeat the experiments in Table 3-4 using the Extra Tree classifier. In Table 5, we only report the F1 score for positive cases due to space limitation.

The results show that the Extra Tree classifier trained on the data conditioning on covariates with higher confounder momentum consistently outperform that conditioning on covariates with lower confounder momentum. This provides us additional supporting evidence for the effectiveness of the identified confounders in cyberbullying detection.

5.2 Qualitative study

Top covariates Top covariates are important because they take main responsibility for predicting cyberbullying behavior. Top covariates in a de-confounded model can causally interpret cyberbullying behavior and enhance the trust in a predictive model [27]. Here we compare the top covariates identified by Extra Tree algorithm with/without de-confounding. In this experiment, we set $K = 3$. The results are listed in Table 6.

We highlight words in different categories with different colors. We can observe that model without de-confounding (Group) identifies online *Informal* words as the most important covariates while de-confounded models detect *Drives, Affective processes, Biological processes* as the top important covariates. Interestingly, studies in psychology have revealed that cyberbullying is an imbalance of power, an urgent, basic, or instinctual need, i.e., biological drives [28]. [32] also shows that overweight and obesity, i.e., biological processes, are the key factors for school-aged children to become victims of bullying. These findings from interdisciplinary studies further corroborate that the proposed de-confounding algorithm is

Table 7: Ten Simpson's pairs in each dataset

Formspring	Twitter
(Tone, differ)	(affect, reward)
(anx, female)	(sad, achieve)
(posemo, differ)	(cogproc, female)
(swear, netspeak)	(bio, feel)
(home, negemo)	(body, sexual)
(negemo, affect)	(health, female)
(insight, Tone)	(health, body)
(tentat, affect)	(drives, Clout)
(death, negemo)	(informal, female)
(sexual, risk)	(swear, female)

effective at identifying plausible causes of cyberbullying behavior. Our algorithm shows the potential to facilitate interdisciplinary collaborations to address the societal challenge of cyberbullying.

Simpson's pairs We also provide some qualitative analyses to gain a deeper understanding of Simpson's paradox in cyberbullying detection. Table 7 includes five Simpson's pairs (potential confounder, main covariate) in both datasets. Here, *affect* represents the affective processes, *differ* represents differentiation such as *but, else*, *anx* is the abbreviation of anxiety, *cogproc* denotes cognitive process, *bio* is the abbreviation of biological process, *swear* includes swear words such as fuck, shit.

Interestingly, Table 7 reveals some counterintuitive findings. For example, studies in public health has concluded that cyberbullying is common among college women [30], words related to female should be a positive predictor of cyberbullying as such. However, our study shows that this conclusion might not be true without controlling for confounders. Specifically, for the Simpson's pair (swear, female), the trend between the *female* and cyberbullying is positive at the aggregate level while it becomes negative when conditioning on the number of *swear* words in the text. Another counterintuitive example is the pair (affect, reward). Intuitively, the feature *reward* is negatively related to cyberbullying while experimental results reveal that the trend reverses when social media users going through different *affective* processes,

6 RELATED WORK

Cyberbullying Detection Existing work of cyberbullying within computer science have primarily focused on building high-accuracy cyberbullying classifiers based on feature engineering. For example, there is a large body of work on developing textual features for detecting cyberbullying behavior [4, 11, 23, 24, 26, 31, 36]. Some work measures the content as the number of offensive terms and studies the changes of words and acronyms used in cyberbullying [24, 31]. Dinakar et al. [8] concatenated TF-IDF features, POS tags of frequent bigrams, and profane words to detect cyberbullying behaviors. Xu et al. [34] presented several off-the-shelf tools such as Bag-of-Words models and LSA- and LDA-based representation learning to predict bullying traces in Twitter. Dani et al. [7] proposed the model to incorporate sentiment information into content features. Their goal was to facilitate cyberbullying detection by capturing the sentiment consistency of normal and bullying posts. In [36], the authors took advantages of visual cues such as features extracted from images and videos to augment the accuracy of cyberbullying detection. Cheng et al. [6] leveraged the multi-modal context of social media platforms and learnt embeddings for social media sessions using heterogeneous network embedding models. In [5], the authors seek to examine cyberbullying detection based on temporal analysis of a corpus of Instagram sessions. Despite its importance, little work in computer science studied the causal interpretation and robustness of a cyberbullying classifier. A similar problem is discussed in [35]. The authors formulate the problem as a sequential hypothesis testing problem and add text-based features based on the feature scores.

Interpretable Machine Learning Our work is also related to interpretable machine learning. The volume of research in interpretability has been quickly growing along with the dominance of Deep Neural Networks. A straightforward approach is to use interpretable models such as linear regression, logistic regression and decision trees to fit data. Then features with large coefficients play key roles in interpreting the predictions. However, the forms of real-world data are usually too complicated to be modeled with simple model class. To explain the predictions of a complex machine learning model, LIME [27] tries to fit local, interpretable models that can explain single predictions of any black-box machine learning model. Koh and Liang [13] attempt to identify training points most related to a given prediction. Specifically, they use influence functions to trace a model's prediction through the learning algorithm and back to its training data. Both methods belong to post-hoc interpretability. Similar work in ad-hoc interpretability studies the feature's importance by permuting the feature's value. Based on the idea introduced in Random Forests [3], Fisher et.al [10] proposed to split the dataset in half and exchange the feature values of the two halves instead of permuting the original features. The knockoff procedure proposed in [2] aims to find truly correlated features based on the "knockoff" variables that are not associated with response but can mirror the structure of the original features.

In contrast to previous work, in this paper, we seek for covariates that are causally related to cyberbullying via blocking the influence of detected p-confounders. This is a critical issue to address because on one hand, a robust cyberbullying classifier can alleviate the problem of lacking in training data, help reach more generalizable conclusions, and on the other hand, understanding the causes of cyberbullying behavior helps improve our ability to predict and to facilitate effective strategies to proactively mitigate its impact.

7 CONCLUSION AND FUTURE WORK

We study a novel problem of robust cyberbullying detection with causal interpretation, and propose an efficient and effective algorithm to identify and block p-confounders from observational data. We first detect potential confounders using Simpson's paradox, and then identify the *most likely* confounders via a data-driven approach. Experimental results show that by controlling confounding bias, the de-confoudnded classifiers can more accurately detect cyberbullying behavior and identify covariates that are causally related to the outcome.

Our work opens several future directions. Firstly, approaches to block the influence of confounders are worth further investigation

as the current data stratification strategy sacrifices the size of training dataset. Secondly, our work assumes all the confounders are observed from data. Therefore, it may fail to consider the uncertainty from the unobserved confounders. Finally, future work should be aimed at better integrating interdisciplinary empirical findings– such as those from psychology and related social sciences–into computational models that detect cyberbullying. Interdisciplinary synergies hold particular promise for identifying, addressing, and preventing this major social problem.

ACKNOWLEDGEMENT

We thank our colleagues from the School of Mathematical and Natural Sciences and School of Social and Behavioral Sciences at ASU, who provided insight and expertise that greatly assisted the research. We would also like to show our gratitude to the 3 "anonymous" reviewers for their so-called insights.

REFERENCES

[1] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Can you Trust the Trend?: Discovering Simpson's Paradoxes in Social Data. In *WSDM*. ACM, 19–27.
[2] Rina Foygel Barber and Emmanuel J Candes. 2016. A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574* (2016).
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[4] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. ACM, 13–22.
[5] Lu Cheng, Ruocheng Guo, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network. In *SDM*.
[6] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 339–347.
[7] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment Informed Cyberbullying Detection in Social Media. In *ECML PKDD*. Springer, 52–67.
[8] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011).
[9] L Christian Elledge, Anne Williford, Aaron J Boulton, Kathryn J DePaolis, Todd D Little, and Christina Salmivalli. 2013. Individual and contextual predictors of cyberbullying: The influence of children's provictim attitudes and teachers' ability to intervene. *Journal of youth and adolescence* 42, 5 (2013), 698–710.
[10] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the" Rashomon" Perspective. *arXiv preprint arXiv:1801.01489* (2018).
[11] Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A unified deep learning architecture for abuse detection. *arXiv preprint arXiv:1802.00385* (2018).
[12] Jaime Roquero Gimenez, Amirata Ghorbani, and James Zou. 2018. Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214* (2018).
[13] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730* (2017).
[14] Manish Mehta, Rakesh Agrawal, and Jorma Rissanen. 1996. SLIQ: A fast scalable classifier for data mining. In *EDBT*. Springer, 18–32.
[15] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.
[16] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. " How Bullying is this Message?": A Psychometric Thermometer for Bullying.. In *COLING*. 695–706.
[17] Sara Pabian, Charlotte JS De Backer, and Heidi Vandebosch. 2015. Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences* 75 (2015), 41–46.
[18] Michael J Paul. 2017. Feature Selection as Causal Inference: Experiments with Text Classification. In *CoNLL*. 163–172.
[19] Judea Pearl. 2000. *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.
[20] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[21] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *ICDMW*. IEEE, 540–547.
[22] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
[23] Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. 2018. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. *arXiv preprint arXiv:1804.03124* (2018).
[24] Elaheh Raisi and Bert Huang. 2017. Co-trained Ensemble Models for Weakly Supervised Cyberbullying Detection. In *NIPS Workshop on Learning with Limited Labeled Data*.
[25] Gunnar Rätsch, Takashi Onoda, and K-R Müller. 2001. Soft margins for AdaBoost. *Machine learning* 42, 3 (2001), 287–320.
[26] Imam Riadi. 2017. Detection of cyberbullying on social media using data mining techniques. *International Journal of Computer Science and Information Security (IJCSIS)* 15, 3 (2017).
[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*. ACM, 1135–1144.
[28] Ken Rigby. 2002. *New perspectives on bullying*. Jessica Kingsley Publishers.
[29] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
[30] Ellen M Selkie, Rajitha Kota, and Megan Moreno. 2016. Cyberbullying behaviors among female college students: witnessing, perpetration, and victimization. *College student journal* 50, 2 (2016), 278–287.
[31] Phoey Lee Teh, Chi-Bin Cheng, and Weng Mun Chee. 2018. Identifying and Categorising Profane Words in Hate Speech. In *ICCDA*. ACM, 65–69.
[32] Musleh uddin Kalar, Tashaba Qaiser Faizi, Maheen Jawed, Sumaira Khalil, Sara M Hussain, Mushkbar Fatima, Faiza Aslam, Hasnain Abbas Dharamshi, and Tahira Naqvi. 2015. Bullying, overweight and physical activity in school children of Karachi. *International Archives of Medicine* 8 (2015).
[33] A. Wald. 1945. Sequential Tests of Statistical Hypotheses. *Ann. Math. Statist.* 16, 2 (06 1945), 117–186. https://doi.org/10.1214/aoms/1177731118
[34] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *NAACL HLT*. Association for Computational Linguistics, 656–666.
[35] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2018. Cyberbullying Detection on Instagram with Optimal Online Feature Selection. In *ASONAM*.
[36] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network.. In *IJCAI*. 3952–3958.

An Empirical Study of Malicious Threads in Security Forums

Joobin Gharibshah
jghar002@ucr.edu
University of California Riverside
Riverside, California

Zhabiz Gharibshah
zgharibshah2017@fau.edu
Florida Atlantic University
Boca Raton, Florida

Evangelos E. Papalexakis
epapalex@cs.ucr.edu
University of California Riverside
Riverside, California

Michalis Faloutsos
michalis@cs.ucr.edu
University of California Riverside
Riverside, California

ABSTRACT

How useful is the information that a security analyst can extract from a security forum? We focus on threads of interest, which we define as: (i) alerts of worrisome events, such as attacks, (ii) offering of malicious services and products, (iii) hacking information to perform malicious acts, and (iv) useful security-related experiences. The analysis of security forums is in its infancy despite several promising recent works. Here, we leverage our earlier work in thread analysis, and ask the question: what kind of information do these malicious threads provide. Specifically, we analyze threads in three dimensions: (a) temporal characteristics, (b) user-centric characteristics (c) content-centric properties. We study threads pulled from three security forums spanning the period 2012-2016. First, we show that 53% of the users asking/selling malicious *Services* on average has 3 posts and initiate 1 thread and 1 day lifetime. Second, we argue that careful analysis can help to identify emerging threats reported in security forums through *Services* and *Alerts* threads and potentially help security analysts prevent attacks. We see this study as a first attempt to argue for the wealth and type of information that can be extracted from security forums.

KEYWORDS

Security forums, Malicious threads, Profiling

ACM Reference Format:
Joobin Gharibshah, Zhabiz Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos. 2019. An Empirical Study of Malicious Threads in Security Forums. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316501

1 INTRODUCTION

Security forums hide a wealth of information, but mining it requires novel methods and tools. The problem is driven by practical forces: there is useful information that could help to improve security, but the volume of the data requires an automated method. The challenge is that there is a lot of "noise", there is lack of structure, and an abundance of informal and hastily written text. At the same time, security analysts need to receive focused and categorized information, which can help their task of shifting through it further. We define the problem more specifically below.

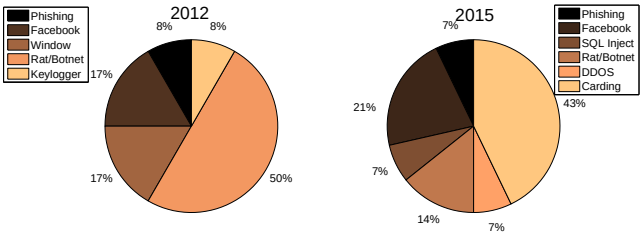


Figure 1: The percentage change of the number of threads discussing different malicious activities in OffensiveCommunity between 2012 and 2015.

Given a security forum, we want to study tempo-behavioral characteristics of **malicious threads**, which we define loosely as threads that a security analyst will want to monitor. We group them in four major categories:

a. *Alerts*: These are threads where users are reporting about being attacked by a hackers or notifying about exploits and vulnerabilities.

b. *Services*: These are threads where users are offering or requesting malicious hacking services or products.

c. *Hacks*: These are threads where users post detailed instructions for performing malicious activities. The difference with the above category is that the information is offered for free here.

d. *Experiences*: These are threads where users share their experience related to general security topics. Often users provide a personal story, a review or an article on a cyber-security concept or event.

There are relatively limited studies on extracting information from security forums, and even less work on using systematic approaches to study the extracted information. We can group prior work in the following categories. First, there is a set of work that focuses on identifying and classifying malicious IP address in security forums.[9–11] Second, there is a group of work that analyzes security forums to identify malicious activity [16, 21]. Third, there are several efforts to detect malicious users [13, 14]. Last but not least, there are some interesting work focusing on emerging threats

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316501

on forums and other social networks [18, 19]. We discuss related work in more detail in our related work section.

In this work, we study malicious threads and users in real forums along three dimensions: (a) temporal, where we analyze the intensity and the evolution of the threads, (b) user-centric, where we characterize user behavior using five features, and (c) content-centric properties. where we analyze the content of the threads by using dominant keywords.

The key contribution of our work is a three-dimensional study of the properties of threads using real data from three security forums: OffensiveCommunity [4], HackThisSite [2] , EthicalHackers [1]. Our data consists of 163k posts and 21k unique threads spanning 2012-2016 across these three forums. Note that we identify the malicious threads by using a method that we developed earlier to: (i) identify malicious threads and, (ii) classify them into these four groups. The development of this method is not part of the contributions of this paper[1] , but we describe it in section 2 for completeness.

We summarize our key findings below.

(a) *Services* providers make short-lived appearances. We find that more than 50% of sellers of malicious services usually make one thread and typically appear in the platform only briefly[2].

(b) Authors of *Hacks* are fairly active users. We find that most of the authors of *Hacks* usually initiate a number of threads and stay active for more than 60 days in the forums which is significantly larger than the average user who posts in malicious threads.

(c) Thread activity has predictive value. We find an indication that analyzing forums carefully can help us anticipate malicious behaviors and even attacks. We found that a *Hacks* on compromising Facebook accounts seemed to have led to an attack reported by *Alerts* thread that discussed a widespread Facebook accounts hacking in 2015, which referred back to the "how to" thread.

(d) The type of malicious activity varies by forum. We find that the type of malicious activity varies by the forum. For example, 45% of malicious threads in OffensiveCommunity are *Hacks*, which is much higher than other forums, while 54% of malicious threads in HackThisSite are *Services*, which exhibits the highest relative activity in this class.

(e) The content of the threads captures interesting trends. We find that the threads in the forum can mirror the trends and concerns of the security community. For example, in 2012, the majority of the discussions focused on malware, including buying and selling and how to exploit botnet and RAT malware, while in the same forum in 2015, the focus was on buying and selling credit card numbers as shown in Figure 1 which is aligned with independently observed trends [3]

2 OVERVIEW OF OUR PRIOR METHOD

In our earlier work, we proposed an approach that identifies and classifies threads of interest from a forum based on one or more sets of keywords that the user provides. We present a high-level description of our approach for completeness, but we do not consider this as part of the contributions of this work.

[1]Currently under submission.

[2]We use the login name of a user to distinguish users. Clearly, the same seller can appear in the platform under different names, which is something that we will investigate in the future.

The approach consists of the following two parts: (a) a similarity-based approach with thread embedding to extract relevant threads of interest, and b) a classification approach based on weighted-embedding method to group relevant threads into user-defined classes.

2.1 Relevant thread extraction

Here we select relevant threads starting from sets of keywords provided by the user. The approach consists of the following phases: (a) a keyword matching step, where we use the user-defined keywords to identify relevant threads that contain these keywords, and (b) a similarity-based phase, where we identify threads that are "similar" to the ones identified above. The similarity is established at the word embedding space.

a. Keyword matching phase. Given a set or sets of keywords, we identify the threads where these keywords appear. A simple text matching approach can distinguish all occurrence of such keywords in the forum threads.

b. Similarity-based phase. Our approach consists of three steps which gets forums, a set of keywords, and set of relevant threads as the input and identify threads of interest as follows:

Step 1. Determining the embedding space. We project every word as a point in a m-dimensional space using a word embedding approach (e.g. Word2Vec). In this way, every word is represented by a vector of m dimensions.

Step 2. Projecting threads. We project all the threads in a multi-dimensional space: both the relevant threads selected from the output of the keyword matching phase and the non-selected ones. The thread projection is a function of the vectors of its words and captures both the average and the maximum values of the vectors of its words[20].

Step 3. Identifying relevant threads. We identify more relevant threads among the non-selected threads that are "sufficiently-close" in terms of similarity to the relevant threads in the thread embedding space.

2.2 Thread Classification

In this part, we classify threads into classes, which are defined by the user. We assume that each class k is defined by a group of words that we denote as $WordClass_k$. In practice these sets of words will be provided by the user and they are inputs to our algorithm.

Our approach can be described as following steps:

Step 1. We create a projection of every class k into the word embedding space by using the words that define the class, $WordClass_k$.

Step 2. We calculate the similarity of every word in the forum v_i for each class k.

Step 3. For each class, we create a weighted embedding by using the similarity to adjust the embedding projection of each word v_i for each class.

Step 4. We use weighted embedding to train an ensemble classifier using supervised learning.

Using the classifier. Given a thread, we calculate its projection in the embedding space, and then we pass it to the classifier to determine its class.

Table 1: The basic statistics of our forums

	OffensiveCommunity	HackThisSite	EthicalHackers
Posts	25538	84125	54176
Threads	3542	8504	8745
Users	5549	5904	2970

3 EXPERIMENTAL RESULTS

We present our experimental results and evaluation of our approach in analyzing extracted malicious threads.

3.1 Our security forums

We use the three forums (OffensiveCommunity, HackThisSite, EthicalHackers) that presented in Table 1. We briefly describe our three forums below.

a. OffensiveCommunity (OC): This forum seems to be on the fringes of legality. As the name suggests, the forum focuses on "offensive security", namely, breaking into systems. Indeed, many posts provide step by step instructions on how to compromise systems, and advertise hacking tools and services.

b. HackThisSite (HT): As the name suggests, this forum has also an attacking orientation. There are threads that describe how to break into websites and systems, but there are also more general discussions about the users' experiences in cyber-security.

c. EthicalHackers (EH): This forum seems to consist mostly of "white hat" hackers, as its name suggests. Many threads are about making systems more secure. However, it contains more *Alerts* and *Experiences* threads than others forums.

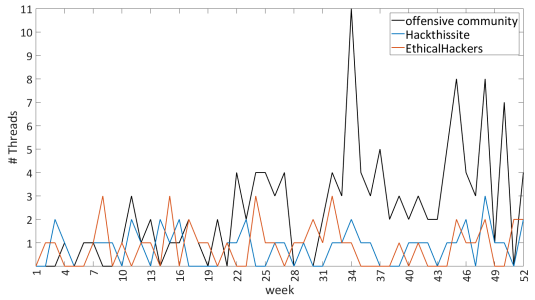


Figure 2: Thread of interest appearing in each week of year 2015 for three forums OffensiveCommunity, HackThisSite, EthicalHackers

Establishing the Groundtruth.

For validating our method, we need groundtruth. To construct the groundtruth, we randomly selected 450 among the relevant threads from each forum as selected by the identification part. The labelling involved three manual evaluations based on the definitions and examples of the four classes, which we listed above. We then combine the "votes", and assign the class selected by the majority. With this process, we labelled 1350 posts in three forums and we show present our labeled data in Table 4.

3.2 Temporal analysis

We study the temporal characteristics of malicious threads. Our goal is to explore forums by looking into information regarding sequence of data over time.

a. The focus of malicious threads varies from year to year. We see in Figure 1 that over two different years the topics of the discussions varied. For example, a topic like "Botnet and Rat" is 50% of the discussions in 2012 but it lost its' popularity in 2015 and possesses only 7% of the discussions. Another topic like "carding" in which users trade stolen credit cards became a major topic in 2015 with about 43% of the discussions of malicious threads in OffensiveCommunity.

b. The malicious activity increased. We find that OffensiveCommunity seems to dominate the other two forums in terms of the absolute number of malicious threads. In 2015 OffensiveCommunity reported malicious threads 2 times more than HackThisSite and EthicalHackers. We can see that during the last five months of 2015 in OffensiveCommunity more malicious threads appear in the forum. In table 2, we see a sample of the discussions happened at week 34 in OffensiveCommunity forums in year 2015.

3.2.1 Case-study: Hacks *activity as an early indication of an attack.* We argue that mining the forums could actually provide information about real events. Here we provide a case where threads of the *Hacks* class were "early indications" of an attack that happened later. Specifically, we identify a link between malicious *Hacks* posted on OffensiveCommunity to hack Facebook accounts in March 2015 to a reported attack by posts of *Alerts* class regarding a widespread Facebook account hacking.

We conducted the following analysis. We plot the time-series of the number of malicious threads in OffensiveCommunity in 2015. We show the time-series in Figure 2. We observe some spikes on these time-series, which we further analyze. One of the spikes was in March 2015, and there is a *Hacks* tutorial thread to break into Facebook accounts with this title "How to Hack Facebook Account with Reconnect Tool". Subsequently, we found other threads in August 2015, in which it was reported that many Facebook's accounts got hacked and the posts argue that attackers used the same techniques as discussed in the earlier threads. The *Alerts* made in a thread with this title: "Hacking Facebook Pages 2015"

We argue that this case-study points to additional layers of functionality that can be built upon our method, that can provide a semi-automated way to extract richer information beyond just reporting malicious threads.

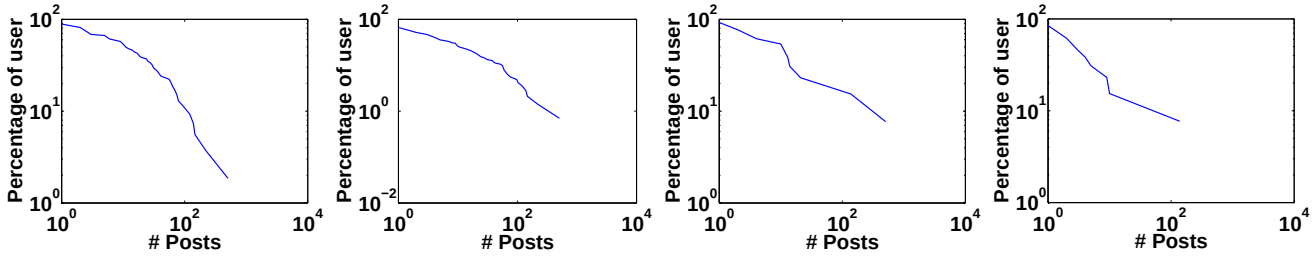
3.3 User-centric analysis

In Figures 3,4,5,6,7, we present the cumulative complementary distribution function of the number of posts per user, the number of threads per user, the number of initiated threads per user, the average length of the posts per users, and the active life time per users for each of the four groups of malicious threads respectively. We focus on the OffensiveCommunity as a forum with indicative behavior in this regard.

(a) *Services* providers make short-lived appearances. We show that more than 50% sellers and buyer of malicious services usually make one thread and they do not post more than 3 times in the thread, also they typically appear in the platform only briefly

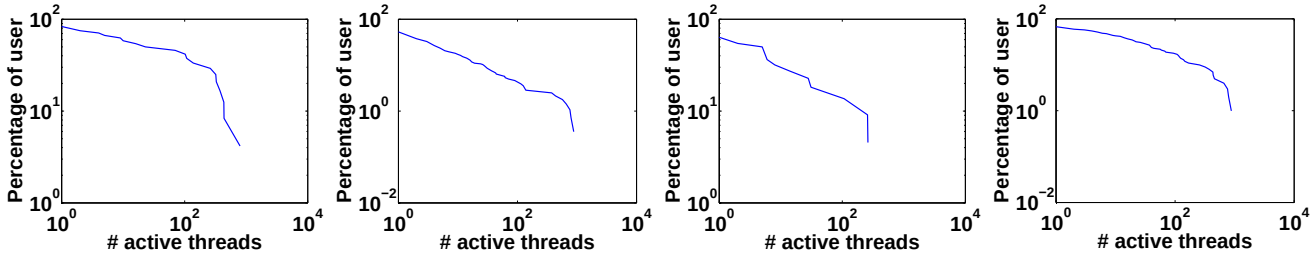
Table 2: Threads of interest in OffensiveCommunity in week 34 in year 2015

Title	Post
Professional Booter/DDoS service OverLoadLab	Hello, welcome to OverLoadLab . We are proud to offer probably the best service for the implementation of DDoS attacks in the world.....
àÿ£ Hacker Wanted MAKE 1000s àÿ£	Digitalmafia is a group of Blackhat hackers from all over the world! Organized Digital Crime Syndicate - ODCS. We are decentralized and can only be contacted through .onion market sites
how to hack facebook account using brute force	What is Brute Force And How to Work Brute force definition Brute force also known as brute force cracking) is a trial and error method used by application programs to decode encrypted data such
PM RIPPER ALERT!!!	Seller with the following IDs - Yahoo ID : XXX - EMail: ... Is a real ripper, i sent him PM and he disappear into thin air... if you love your PM please do not send to this guy. He is here to rip off. Be warned!!!



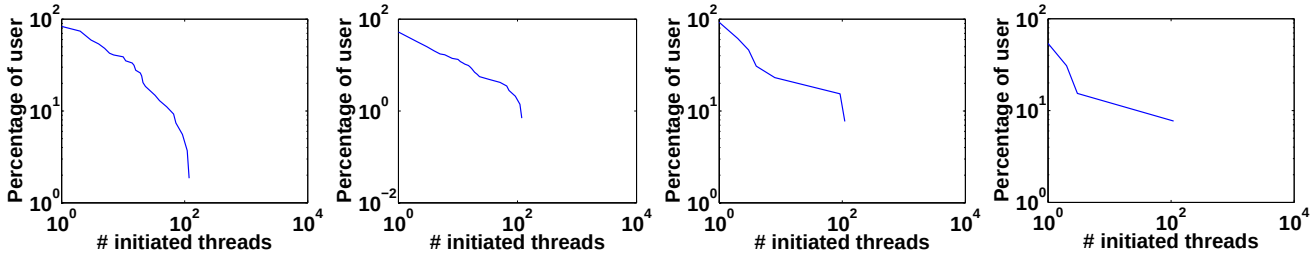
(a) *Hacks* (b) *Services* (c) *Alerts* (d) *Experiences*

Figure 3: CCDF of the number of posts per user who posted malicious threads (log-log scale) in OffensiveCommunity. More than 54% of the users who write in *Services* has less than 3 posts in the forum.



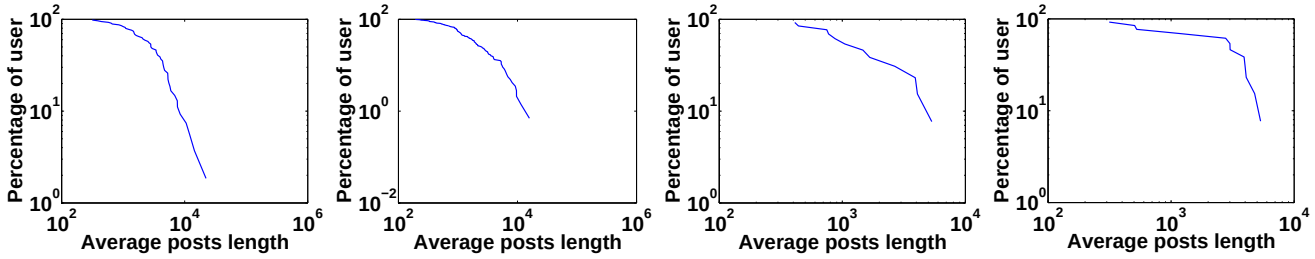
(a) *Hacks* (b) *Services* (c) *Alerts* (c) *Experiences*

Figure 4: CCDF of the number of active threads per user who posted malicious threads (log-log scale) in OffensiveCommunity. More than 63% of the users who write in *Services* were active in only one or two threads.



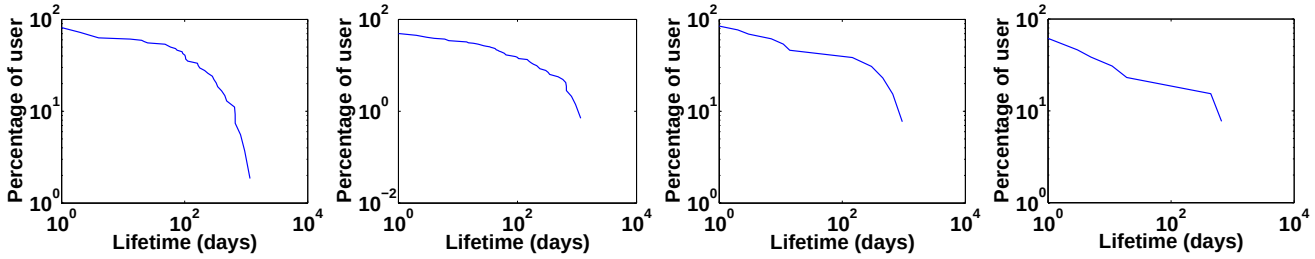
(a) *Hacks* (b) *Services* (c) *Alerts* (d) *Experiences*

Figure 5: CCDF of the number of threads initiated per user who posted malicious threads (log-log scale) in OffensiveCommunity. More than 68% of the users who write in *Services* initiated less than 2 threads.



(a) *Hacks* (b) *Services* (c) *Alerts* (d) *Experiences*

Figure 6: CCDF of the average posts length per user who posted malicious threads (log-log scale) in OffensiveCommunity. More than 50% of the users write posts with average length of 870 characters in *Alerts* threads, while in *Hacks* threads the average post length is less than 4100 characters.



(a) *Hacks* (b) *Services* (c) *Alerts* (d) *Experiences*

Figure 7: CCDF of the average lifetime length per user who posted malicious threads (log-log scale) in OffensiveCommunity. More than 50% of the users who start *Services* threads appear only for one day in the forum.

for at most one day (at least with a given login name). Moreover compared to the users who posting in other categories, here they write shorter posts with an average of 870 characters. This behavior is different than what have seen in other social trading networks where people only do trading. [6].

(b) *Hacks* providers are mostly active users. We find that most of the *Hacks* providers usually initiate larger number of threads and stay active more than 60 days in the forums, which is significantly larger than other users posting in the malicious threads. These users on average use longer posts length compared to users in other categories with more than 4100 characters.

Moreover, we see in all the cases the distributions are skewed, that is, most of the users contribute few posts in the forums and engage with few threads. In OffensiveCommunity, 85% of users who post malicious threads have less than 10 posts in the forum, while 3.8% of such users post more than 50 posts. We find that more than 60% of the users have the lifetime less than 13 days in the forums. It means that for such users the difference between their first and the last post in the forums is less than 13 days. This skewed behavior is typical for online users and communities [7].

3.4 Content-centric analysis

Here, we study the content of malicious threads in forums, although we have done some of that already in our temporal analysis.

a. The type of malicious activity variation by forum. We find that the type of malicious activity varies by the forum. Usually in all forums the *Services* threads have the highest activity. However, in other categories we can see that 45% of malicious threads in OffensiveCommunity is in *Hacks* which is more than other forums. EthicalHackers has the most number of *Alerts* threads compared to other forums. HackThisSite with 54% of the threads in *Services* has the highest activity in this class among other forums. We show more details in Table 4.

Table 3: *WordClass*, the set of words which "define" each class.

Hacks	*Services*	*Alerts*	*Experiences*
tutorial	tool	announced	article
guide	price	reported	story
steps	pay	hacker	challenge

Table 4: The breakdown of the data per class in forums.

	OffensComm.		HackThisSite		EthicalHackers	
Total	450		450		450	
	#	%	#	%	#	%
Hacks	202	45%	49	11%	42	9%
Services	204	45%	242	54%	166	37%
Alerts	27	6%	37	8%	78	17%
Experiences	17	4%	128	28%	164	36%

b. Identifying surprising keywords. We extract the dominant keywords in each class of the malicious threads, and we see words that we did not expect. In more detail, we consider the frequency

(a) *Services* (b) *Hacks* (c) *Alerts* (d) *Experiences*

Figure 8: Word cloud for the different classes of malicious threads for OffensiveCommunity.

of the keywords in each class of malicious threads. We show the results for OffensiveCommunity.

First, interestingly, we found that less than 40% of $WordClass_k$, which we introduced in section 2 and depicted in Table 3 appear as the most dominant keywords in malicious threads classes as we show in Figure 8. For example, the most frequent words in class *Alerts* in OffensiveCommunity forums are: “hacker", “Indian", “Pakistan", “hacked". Note that among these words, we only used the word “hacked" to define class *Alerts* previously. Similarly in *Hacks* class the words “tutorial", “code", “file" dominate in terms of frequency. Among these words, only the word “tutorial" was used to define this class.

Second, we find some surprising frequent words. For example, it is quite surprising that the words “Indian" and "Pakistan" emerge as popular words. For exmaple there is an *Alerts* thread with title “Pakistan Under Big Cyber Attack By Indians - 15 August” with discuss about a cyber-was between two countries of Pakistan and India.

The importance of such analysis is that even though we extracted the malicious threads with an initial set of keywords (Table 3), our analysis can point to interested related concepts that were not anticipating ahead of time. In fact, we intend to create an iterative approach, where we start with an initial set of keywords, and in each iteration we can refine the keywords of interest in order to explore naturally occurring themes.

4 RELATED WORK

We summarize related work in extracting and classifying entities from security forums.

a. Identifying IP addresses in security forums. There is a group of efforts that study IP addresses in security forums to identify and classify malicious IP addresses in such forums [9, 10]. They have proposed transfer learning based approach that can use training information on one forum to develop sufficient training data for a new forum [11]. In another work [8], they focus on the spatiotemporal properties of Canadian IP addresses in forums without employing any identification and classification methods.

b. Identifying entities of interest. A few recent studies identify malicious services and products in security forums by focusing on their availability and price [6, 15, 16] Another interesting work [21] uses a word embedding technique focusing identifying vulnerabilities and exploits.

c. Identifying malicious users. there are several efforts to study the users of security forums, group them into different classes, and identify their roles and social interactions [5, 12–14, 22].

d. Identifying malicious events. Another work [18] identifies emerging threats by monitoring the behavior of malicious users and correlating it with information from security experts on Twitter. Moreover, there is a comprehensive study on the vulnerability reported on twitter. [17] Another study [19] detects emerging security concerns by monitoring the keywords used in forums and other online platforms, such as blogs.

5 CONCLUSION

There is a wealth of information in security forums, but still, the analysis of security forums is in its infancy, despite several promising recent works.

The goal of our work is provide empirical, but concrete indication, of useful information from threads that have been identified as threads of interest. We consider three dimensions of properties: (a) temporal, (b) user-centric, (c) content-centric properties. Using real data from security forums, we show that we can find interesting and often actionable pieces of information from these forums.

In the future, we plan to extend our work in two different dimensions. First, we want to develop automated methods to extract useful and actionable information from such threads in a form that can be readily used by a security analyst. Second, we want to develop visualization techniques and metrics that can easily highlight properties and phenomena of interst.

6 ACKNOWLEDGMENT

This material is based upon work supported by DHS ST Cyber Security (DDoSD) HSHQDC-14-R-B00017 grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding institutions.

REFERENCES

[1] 2018. Ethical hacker Forums. https://www.ethicalhacker.net/forums/
[2] 2018. Hack This Site Forums. https://www.hackthissite.org/forums/
[3] 2018. Nasdaq. https://www.nasdaq.com/article/credit-card-fraud-and-id-theft-statistics-cm520388
[4] 2018. Offensive Community Forums. http://www.offensivecommunity.net
[5] A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen. 2014. Descriptive Analytics: Examining Expert Hackers in Web Forums. In *2014 IEEE Joint Intelligence and Security Informatics Conference*. 56–63. https://doi.org/10.1109/JISIC.2014.18

[6] Luca Allodi. 2017. Economic factors of vulnerability trade and exploitation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1483–1499.

[7] P. Devineni, D. Koutra, M. Faloutsos, and C. Faloutsos. 2015. If walls could talk: Patterns and anomalies in Facebook wallposts. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 367–374. https://doi.org/10.1145/2808797.2808880

[8] R. Frank et al. 2016. Location, Location, Location: Mapping Potential Canadian Targets in Online Hacker Discussion Forums *(EISIC '16)*.

[9] Joobin Gharibshah et al. 2018. Mining actionable information from security forums: the case of malicious IP addresses. *CoRR* abs/1804.04800 (2018). arXiv:1804.04800

[10] Joobin Gharibshah, Tai Ching Li, Maria Solanas Vanrell, Andre Castro, Konstantinos Pelechrinis, Evangelos E. Papalexakis, and Michalis Faloutsos. 2017. InferIP: Extracting Actionable Information from Security Discussion Forums. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17)*. ACM, New York, NY, USA, 301–304. https://doi.org/10.1145/3110025.3110055

[11] Joobin Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos. 2018. RIPEx: Extracting Malicious IP Addresses from Security Forums Using Cross-Forum Learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'18)*. Springer International Publishing, Cham, 517–529.

[12] Thomas J Holt, Deborah Strumsky, Olga Smirnova, and Max Kilger. 2012. Examining the Social Networks of Malware Writers and Hackers. *International Journal of Cyber Criminology* 6 (01 2012), 891–903.

[13] W. Li and H. Chen. 2014. Identifying Top Sellers In Underground Economy Using Deep Learning-Based Sentiment Analysis. In *2014 IEEE Joint Intelligence and Security Informatics Conference*. 64–67.

[14] E. Marin, J. Shakarian, and P. Shakarian. 2018. Mining Key-Hackers on Darkweb Forums. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*. 73–80.

[15] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. 2011. An Analysis of Underground Forums. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC '11)*. ACM, New York, NY, USA, 71–80.

[16] Rebecca S. Portnoff, Sadia Afroz, Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Tools for Automated Analysis of Cybercriminal Markets. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 657–666.

[17] Carl Sabottke, Octavian Suciu, and Tudor Dumitras. 2015. Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits. In *24th USENIX Security Symposium (USENIX Security 15)*. USENIX Association, Washington, D.C., 1041–1056.

[18] Anna Sapienza, Alessandro Bessi, Saranya Damodaran, Paulo Shakarian, Kristina Lerman, and Emilio Ferrara. 2017. Early warnings of cyber threats in online discussions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 667–674.

[19] Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining Online Chatter for Emerging Cyber Threats. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 983–990. https://doi.org/10.1145/3184558.3191528

[20] Dinghan Shen et al. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *ACL 2018*.

[21] Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. DarkEmbed: Exploit Prediction With Neural Language Models. In *IAAI2018*.

[22] Xiong Zhang, Alex Tsang, Wei T. Yue, and Michael Chau. 2015. The Classification of Hackers by Knowledge Exchange Behaviors. *Information Systems Frontiers* 17, 6 (Dec. 2015), 1239–1251.

Understanding Cyberbullying on Instagram and Ask.fm via Social Role Detection

Hsien-Te Kao
University of Southern California, Information Sciences Institute
hsientek@usc.edu

Shen Yan
University of Southern California, Information Sciences Institute
shenyan@usc.edu

Di Huang
University of Southern California, Information Sciences Institute
dh_599@usc.edu

Nathan Bartley
University of Southern California, Information Sciences Institute
nbartley@usc.edu

Homa Hosseinmardi
University of Southern California, Information Sciences Institute
homahoss@isi.edu

Emilio Ferrara
University of Southern California, Information Sciences Institute
ferrarae@isi.edu

ABSTRACT

Cyberbullying is a major issue on online social platforms, and can have prolonged negative psychological impact on both the bullies and their targets. Users can be characterized by their involvement in cyberbullying according to different social roles including victim, bully, and victim supporter. In this work, we propose a social role detection framework to understand cyberbullying on online social platforms, and select a dataset that contains users' records on both *Instagram* and *Ask.fm* as a case study. We refine the traditional victim-bully framework by constructing a victim-bully-supporter network on Instagram. These social roles are automatically identified via ego comment networks and linguistic cues of comments. Additionally, we analyze the consistency of users' social role within *Instagram* and compare users' behaviors on *Ask.fm*. Our analysis reveals the inconsistency of social roles both within and across platforms, which suggests social roles in cyberbullying are not invariant by conversation, person, or social platform.

CCS CONCEPTS

• **Networks → Online social networks**; • **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

cyberbullying, social role, social networks

ACM Reference Format:
Hsien-Te Kao, Shen Yan, Di Huang, Nathan Bartley, Homa Hosseinmardi, and Emilio Ferrara. 2019. Understanding Cyberbullying on Instagram and Ask.fm via Social Role Detection. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316505

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316505

1 INTRODUCTION

Cyberbullying is a major problem among teenagers on online social platforms. One in four middle and high school students reported being bullied at least once a week between 2015 and 2016 [6]. These students experienced both emotional (e.g., teasing, inappropriate comments) and social bullying (e.g., spreading rumors, embarrassing comments) on social media. Students who experienced cyberbullying are more likely to develop depression, anxiety, sleep difficulties, have a lower academic achievement, and drop out of school [3]. The bullies suffer as well: students who cyberbullied other students are at increased risk for mental health, behavioral problems, academic problems, and substance use [5]. Cyberbullying has prolonged negative impacts on both bullies and their targets, thus early detection is essential to minimize the potential harms.

There is extensive research on understanding individual behaviors and interactions during cyberbullying. Wegge *et al.* [19] showed victims tend to be cyberbullied by the same perpetrator who bully them offline. Festl and Quandt[7] found perpetrators in offline bullying are likely to bully other online users, whereas victims in offline bullying may not be victims in cyberbullying. Bastiaensens *et al.* [2] discovered bystanders will help the victims if the incident is severe, but they will also bully the victims when they are good friends with the perpetrator. These studies viewed cyberbullying as individual incidents rather than patterns of behaviors that might carry over between different incidents, different avenues of communication within a platform, and different platforms altogether. In addition, these studies focused heavily on victim-bully interactions that oversimplify cyberbullying interactions when there are other roles involved, like victim supporter.

To better understand individual behaviors on online social platforms, user interactions can be expressed by the networks and different interaction patterns can be decomposed into different social roles. Social role detection has been used to classify different types of user interactions. Akar and Mardikyan [1] reviewed various identified user roles on social platforms including Wikipedia, Reddit, and Twitter. Buntain and Golbeck [4] discovered the "answer-person" and "discussion-person" roles on Reddit and the consistency of these social roles across sub-communities of Reddit. Lumbreras *et al.* [13] developed a social role detection algorithm for coexistence roles in the community based on growth models for trees.

In this work, we aim to utilize social role detection to understand cyberbullying on a complex social network. We select a dataset [9]

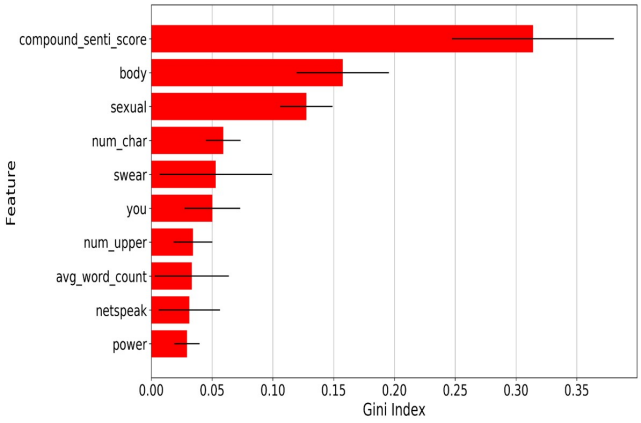


Figure 1: Top 10 important features.

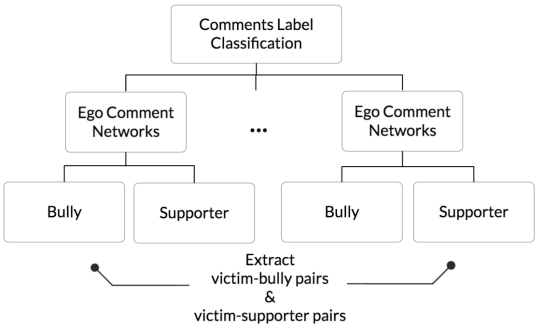


Figure 2: Pipeline of social role detection method.

that consists of users on such Instagram and Ask.fm to understand cyberbullying on online social platforms. Both platforms were ranked among the top five cyberbullying platforms in 2013 [18], and recent reporting shows Instagram continues to be a top platform [12]. In addition to the traditional victim-bully framework in cyberbullying, we further introduce the victim-supporter role based on bullying literature [17]. There are two types of supporting behavior, positive supporting comments that side with the victim and negative aggression toward the bullies, and our work focuses on the latter case. We propose a social role detection framework for victim-bully-supporter network that answers the following three questions: *(i)* whether the victim, bully, and supporter social roles exist on Instagram and Ask.fm, (ii) if it is possible to identify these roles in an automated fashion, and, (iii) whether a user's social roles and behaviors are consistent within and across online social platforms.

2 RELATED WORK

2.1 Social role detection

There are numerous studies about role detection on social platforms, such as Reddit. Buntain *et al.* [4] set out to identify the "answer-person" social role on Reddit in an automated fashion using Decision Trees for classification. To do this, they constructed ego-networks of users on a collection of subreddits and hand-labeled those users who conformed to the "answer-person" role they were looking for. Then, they used various network metrics as features for their classifier, whose performance suggested that network features may be very predictive for this specific role. The presented work differs in that it makes use of both linguistic and network features to examine multiple roles.

Similarly, Kou *et al.* [11] set out to study knowledge production about user experience work in the specific subreddit "/r/userexperience". They used a mixed methods approach consisting of statistical tests and qualitative content analysis to analyze the roles of the top users by activity in their dataset. This role analysis resulted in five distinct roles that gave insight into how learning and knowledge distribution happens within a community of user experience practitioners. The presented work differs methodologically in that roles are fixed and analyzed in a more automated fashion.

Another role detection study was done by Akar *et al.* [1] on the Turkish forum *incisozluk.com.tr*. The authors collected the data, constructed bipartite graph representations of the users and the topics they posted in, and projected that to look only at the users network. They then used the network properties of each member and a developed questionnaire to describe the different structural roles they discovered in the data. Importantly, they discovered that certain behaviors and perceptions seemed to be moderated by the social role of the user: for instance, perceived critical mass of a topic is a more prominent factor for visitors and passive members of a community, but not as much for content generators.

2.2 Cyberbullying detection

Detection of cyberbullying is a difficult problem that is hard to define as it manifests in different ways on different online social networks. Hosseinmardi *et al.* [9] acknowledged this and collected their dataset consisting of common users between Instagram and Ask.fm to characterize how much of an effect different platforms might have on how cyberbullying manifests. They performed various statistical tests on the number of positive and negative words found in posts across both platforms, as well as analysis of network-related features in their data. For instance, they found correlations between being positive on one network and being positive on the other network, and similarly for being negative. However, given limitations of their methods they do not label comments as bullying or not nor do they explore social roles contained in their data. Zhao *et al.* [21] proposed an automatic representation learning method for cyberbullying detection named embeddings enhanced Bag-of-Words model (EBoW). EBoW combines traditional linguistic features and bullying features based on insulting words with word embeddings. However, their detection method only considers the language properties but cannot capture the interactions between users.

3 DATASET

We use a dataset [8, 9] that contains user records on both Instagram and Ask.fm collected from August 2011 to June 2014. It consists of 14,063 users' information on Instagram and 8,696 of these users'

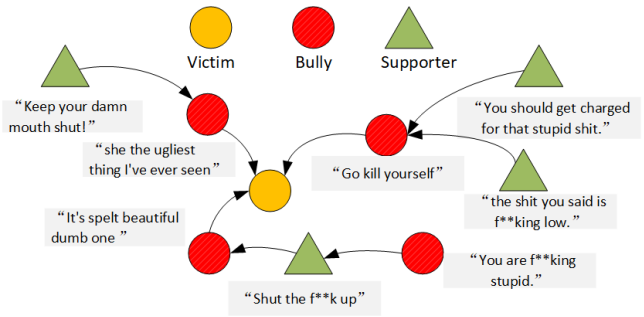


Figure 3: Example of aggressive comments in a Instagram media session. Yellow – victim, red – bully, green – supporter.

Ask.fm information who also have an Ask.fm account. The Instagram data include 1,738,850 media sessions and 6,816,844 comments. These media sessions are in some form of media (e.g., video, picture) posted by the user to the user's account. The user has the opportunity to supply a caption to their own media when they submit the media, at which point it is either presented to the user's followers or the broader public if it is a public account. This allows for other users to comment on the media session. The Ask.fm data include 11,047,124 question-answer pairs. A user can send a question anonymously or non-anonymously to another user profile on Ask.fm. We separate the dataset into two sets of unique questions, one with 3,144,000 non-anonymous questions and one with 5,702,324 anonymous questions, and each question receives 1.7 answers on average.

4 AGGRESSIVE COMMENT CLASSIFIER

We labeled 13,310 Instagram comments as either aggressive or not in order to train a classifier to label the Instagram comments. We divided the comments into four different sets randomly, each of which was labeled by one of the four members of our team using the same set of instructions.

Each comment is given either a label 1 as aggressive or label 0 as not aggressive. In the end, there are about 12% comments labeled as aggressive and the rest are labeled as not aggressive. The labeled Instagram comments are chosen from top 200 media sessions ranked by the number of insulting words which appeared in the comments of the media session, the percentage of negative comments, and the absolute number of negative comments on the media session. The original insulting words were drawn from a aggressive words dictionary[1]. We expanded this dictionary with functionally similar terms from our corpus using the Gensim implementation of word2vec [14, 16]. We have a total of 739 insulting words after expansion. After this, we extract 83 linguistic features from each comment: 73 features from LIWC2015 categories [15], the compound sentiment score from VADER [10], and 7 basic linguistic features (e.g., the number of upper case words). We use LIWC2015 categories to provide a higher-level linguistic pattern extraction, the VADER compound sentiment score because it considers emojis and punctuation, and basic linguistic features to capture linguistic difference between comments. Considering the class imbalance and the feature properties, we train a binary random forest classifier with non-linear HSIC (Hilbert-Schmidt Independence Criterion) Lasso feature selection [20] and parameter tuning to label the remaining comments.

[1]https://www.noswearing.com/dictionary

Table 1: Model performance for minority class (aggressive comment) with class imbalance 88-12 for nested 10-fold cross-validation.

Precision	Recall	F1 Score	False Positive
0.40	0.35	0.37	0.04

In Figure 1, we can see that our binary random forest classifier is labeling the Instagram comment as aggressive or not based on sentiment score, linguistic properties, and context. Compound sentiment score is the strongest feature because aggressive comments have higher negative sentiment score compared to non-aggressive comments. The classifier is labeling based on sexual, body, swear, and power words that target someone in addition to large sentence with a lot of upper case words and internet language. The selected word categories are often used in cyberbullying, and long sentence and upper case word are signs for strong negative emotion. In Table 1, we show the binary random forest classifier performance with accuracy 0.89. We apply the trained classification model to label all 6.8 million Instagram comments.

5 SOCIAL ROLE DETECTION

In this work, we focus our analysis on three different social roles in cyberbullying: victim, bully, and supporter. Figure 2 describes the pipeline of our social role detection method. Social roles are identified based on the aggressive label of comments and each user's ego comments networks. Figure 3 shows a typical example of a real bullying event in our dataset. Bullies posted aggressive comments to the author of the media session (victim), and supporters tried to defend the victim by attacking the bullies. Based on the observations above, we define users' social roles by their interaction patterns:

- **Victim:** users who receive aggressive comments repeatedly.
- **Bully:** users who send aggressive comments repeatedly.
- **Supporter:** users who send aggressive comments to bullies.

5.1 Ego network

In order to identify victim-bully and victim-supporter pairs, we build ego comment networks for each user. We only include the aggressive comments we classified from Section §4. An ego comment network $G_i(V_i, E_i, W_i)$ for user i is built based on all aggressive comments of i's media sessions (Figure 5 shows an ego comment network example for a particular user). V_i are all the users who commented on i's media sessions, including i. E_i represent the set of all edges e_{jk} between any user j and k, when there at least one aggressive comment from user j toward user k, and $w_{jk} \in W_i$ indicates number of media sessions with aggressive comments from user j to k, e.g., the edge $e_{jk} : (j \rightarrow k, 3)$ means that j has sent aggressive comments to i in 3 media sessions. In order to capture the recurrent nature of cyberbullying behaviors, any user j who has sent aggressive comments in more than one media sessions

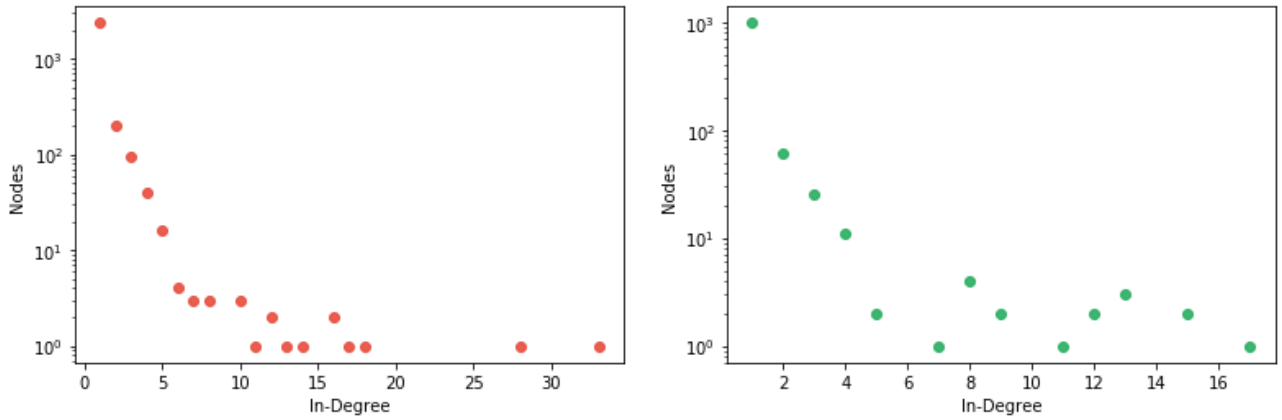


Figure 4: Number of nodes as function of in-degree for victim-bully/victim-supporter pair networks.

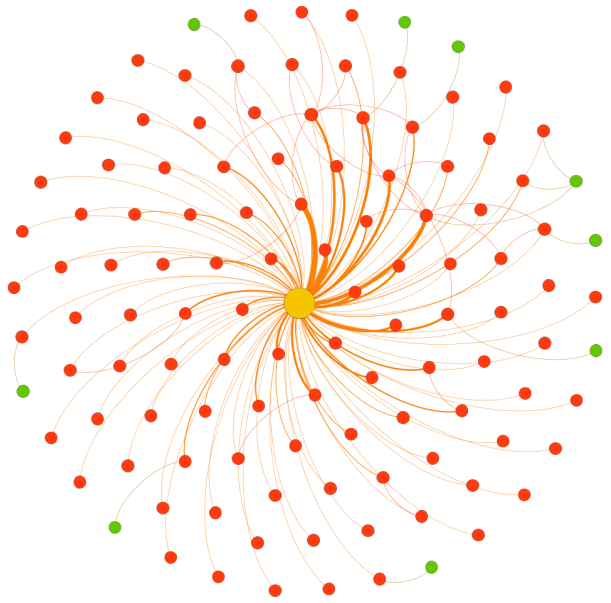


Figure 5: Example of ego comment network. yellow – victim, red – bully, green – supporter.

$(w_{j,i} > 1)$ towards user i, is considered as potential bully. The supporters are the users who have sent aggressive comments to the bullies identified above.

After filtering out the one-time aggressive behaviors, we identify 1,800 victim-bully pairs and 712 victim-supporter pairs from Instagram networks. Figure 4 shows the in-degree distribution of the victim-bully pair network, where in-degree values represent the number of bullies the user has. Comparing with the in-degree distribution in Figure 6, the number of bullies significantly decreased, which indicates that most bullying behaviors are temporal and occasional. Based on the victim-bully pair network, 13 users are bullied by more than 10 people, and the maximum number of bullies is 33. While in terms of the out-degree (i.e., number of users that he/she has bullied), only 2 users bullied more than one people repeatedly. Similar patterns also appear in the victim-supporter pair networks. Most users only has one supporter, and the users who has been identified as supporter only support one specific user. Only 9 users have more than 10 supporters.

5.2 Aggregated user network

Based on each ego bullying network, we assign users to one of the three social roles: victim, bully, or supporter. However, those individual-based social roles may differ across different interactions. In order to analyze a user's role consistency, we create an weighted aggregate victim-bully-supporter network $G(V, E, W)$ based on all aggressive comments on Instagram. Here, each node represents a user involved in Instagram media sessions. A directed edge e_{ij} from user i to user j represents there is at least one aggressive comment sent by i to j. Furthermore, we weight the directed link from i to j by counting the number of aggressive comments made by i towards j (i.e., w_{ij}). The case would either be that j was the owner of the media session where i commented or j replies to i in other users' media sessions.

Consequently, we generate a graph consisting of 441,523 nodes and 466,309 edges. Figure 6 shows the in-degree/out-degree distribution of the victim-bully-supporter network. The maximum in-degree value is 8,445, which means a user received aggressive comments from 8,445 users. The maximum out-degree value is 487, which means a user has sent aggressive comments to 487 users. For node i, if its in-degree d_i^{in} is much larger than its out-degree d_i^{out}, i is a global victim. When both d_i^{in} and the sum of the weights of incoming edges are larger, the bullying behaviors towards user i is more severe. On the other hand, if out-degree d_i^{out} is much larger than in-degree d_i^{in}, then user i is regarded as a global bully. Moreover, i could be a hater who bullied numerous users when

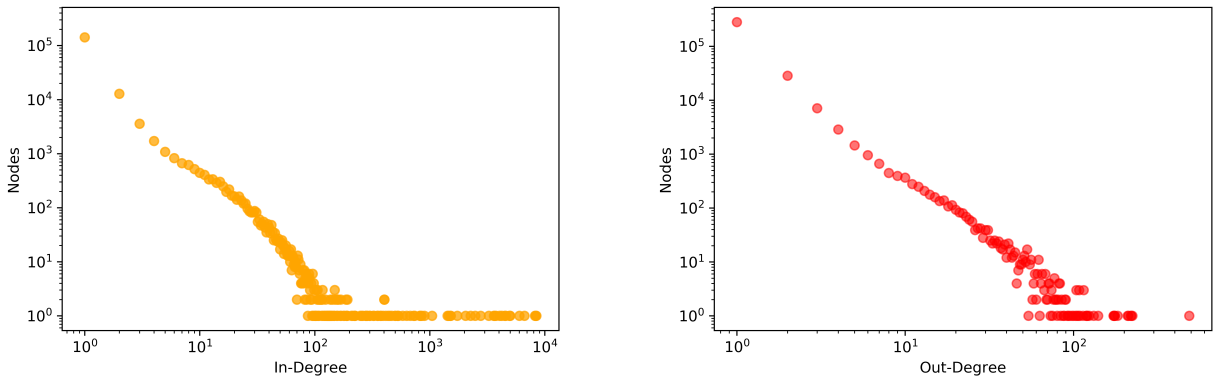


Figure 6: Number of nodes as function of in-degree/out-degree for victim-bully-supporter network. The in-degree value represents the number of bullies the user received; the out-degree value indicates the number of users he/she bullied before.

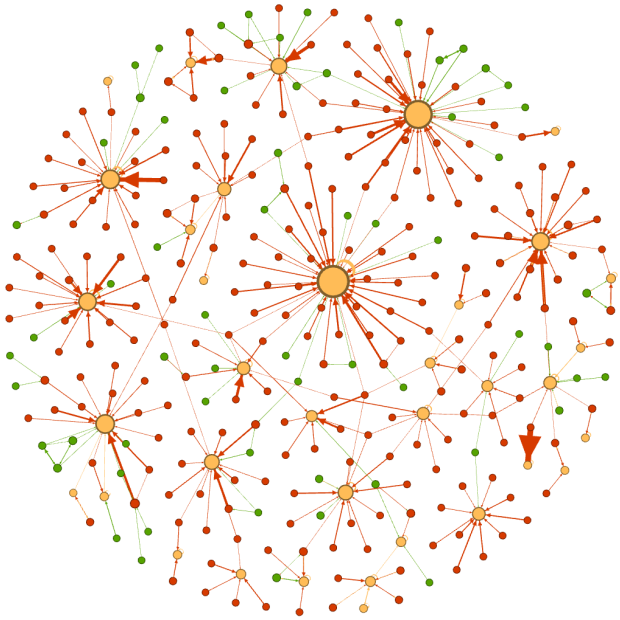


Figure 7: Victim-bully-supporter network (subgraph). Yellow – victim, red – bully, green – supporter.

d_i^{out} is relatively large. In addition, i is a global supporter only when his/her links were directed to global bullies.

By comparing the global social role and individual-based social roles for each user, we find that the three social roles have different consistency across media sessions. Figure 7 is a subgraph based on the largest component of the victim-bully-supporter network. The role assigned to a node is the majority of its individual-based roles through different media sessions. As Figure 7 shows, victims and bullies are more consistent between the roles identified based on ego networks and the aggregate network. However, some supporters also bullied victims on the aggregate network, which shows the inconsistency of the social role supporter across media sessions.

5.3 Consistency in Ask.fm

We rank the users on Instagram by the total number of aggressive comments that they posted on Instagram and look at these top-ranked users' Ask.fm data to examine how consistent the roles are between platforms. We select all the aggressive comments on Instagram that these users were involved in and identify the original author, the commenter, and anyone who is mentioned in the comment. From the top 1,000 ranked Instagram users, we identify 27 bully-victim pairs where the top-ranked user is the bully and the victim was either mentioned or had interaction with the bully on both platforms (61 pairs are present when we look at all the users). Interestingly, none of the roles in these pairs seemed to be consistent, as the identified bullies were always responding to their own media sessions on Instagram, suggesting they were inviting or expecting comments in their media sessions. The corresponding question/answer pairs on Ask.fm for these bully-victim pairs were all benign. The language in some of Instagram comments seems to suggest a playful aggression.

6 CONCLUSIONS AND FUTURE WORK

Cyberbullying has been growing on online social platforms and generates massive negative user experiences in addition to causing prolonged negative psychological impacts on both bully and their targets. Numerous researchers have studied cyberbullying in specific contexts (e.g., conversation, topic, relationship, social platform) using a constrained victim-bully framework. We here proposed a fine-grained social role detection method automatically discovering victim-bully-supporter interactions on online social platforms to understand cyberbullying. We chose a dataset that has the records from common users on both Instagram and Ask.fm as our case study. Comparing the consistency of users' social roles within Instagram and users' behaviors on Ask.fm, our work revealed that bullies and victims are more likely to have consistent social roles within Instagram, while supporters show more complex interaction patterns. Users' behavior patterns on Ask.fm also deviate from their social roles identified in Instagram. These observations show the consistency of social roles is specific to a conversion topic or

person. Therefore, it is difficult to classify users' behaviors solely based on their social role in cyberbullying detected in a specific conversation, relationship, or social platform. This new insight is essential in designing safer online social ecosystems with adaptive cyberbullying social role detection.

This work mainly focused on the detection of pre-defined social roles in cyberbullying scenarios. For future work, we are considering to extend our methods to the discovering different behavior patterns as their social roles. We plan to leverage entity recognition and relation extraction techniques to discover the relationships between users from the comments' content, to enable capturing more complex behavior patterns and give us more insights into cyberbullying events.

REFERENCES

[1] Ezgi Akar and Sona Mardikyan. 2018. User Roles and Contribution Patterns in Online Communities: A Managerial Perspective. *SAGE Open* 8, 3 (2018), 2158244018794773.

[2] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann Desmet, and Ilse De Bourdeaudhuij. 2014. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior* 31 (2014), 259–271.

[3] Sheri Bauman, Russell B Toomey, and Jenny L Walker. 2013. Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence* 36, 2 (2013), 341–350.

[4] Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 615–620.

[5] Centers for Disease Control and Prevention. 2017. Youth Risk Behavior Surveillance. https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2017/ss6708.pdf. (2017).

[6] Melissa Diliberti, Michael Jackson, and Jana Kemp. 2017. Crime, Violence, Discipline, and Safety in US Public Schools: Findings from the School Survey on Crime and Safety: 2015-16. First Look. NCES 2017-122. *National Center for Education Statistics* (2017).

[7] Ruth Festl and Thorsten Quandt. 2013. Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet. *Human communication research* 39, 1 (2013), 101–126.

[8] Homa Hosseinmardi, Amir Ghasemianlangroodi, Richard Han, Qin Lv, and Shivakant Mishra. 2014. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 244–252.

[9] Homa Hosseinmardi, Shaosong Li, Zhili Yang, Qin Lv, Richard Han, Rahat Ibn Rafiq, and Shivakant Mishra. 2014. A comparison of common users across instagram and ask. fm to better understand cyberbullying. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. IEEE, 355–362.

[10] C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*. International Conference on Weblogs and Social Media, 216–225.

[11] Yubo Kou, Colin Gray, Austin Toombs, and Robin Adams. 2018. Knowledge Production and Social Roles in an Online Community of Emerging Occupation: A Study of User Experience Practitioners on Reddit. (2018).

[12] Taylor Lorenz. 2018. "Teens Are Being Bullied 'Constantly' on Instagram". *The Atlantic* (10 2018).

[13] Alberto Lumbreras, Bertrand Jouve, Julien Velcin, and Marie Guégan. 2017. Role detection in online forums based on growth models for trees. *Social Network Analysis and Mining* 7, 1 (2017), 49.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015).

[16] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[17] Stopbullying. 2017. The Roles Kids Play in Bullying. https://www.stopbullying.gov/what-is-bullying/roles-kids-play/index.html. (2017).

[18] Ditch the Label. 2013. The Annual Cyberbullying Survey. (2013).

[19] Denis Wegge, Heidi Vandebosch, and Steven Eggermont. 2014. Who bullies whom online: A social network analysis of cyberbullying in a school context.

[20] Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P Xing, and Masashi Sugiyama. 2014. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation* 26, 1 (2014), 185–207.

[21] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*. ACM, 43.

Domain Generation Algorithms detection through deep neural network and ensemble

Shuaiji Li
DiDi Research America
Mountain View, CA, USA
sl6486@nyu.edu

Tao Huang
DiDi Research America
Mountain View, CA, USA
taohuang@didichuxing.com

Zhiwei Qin
DiDi Research America
Mountain View, CA, USA
qinzhiwei@didichuxing.com

Fanfang Zhang
DiDi Research America
Mountain View, CA, USA
zhangfangfangff@didichuxing.com

Yinhong Chang
DiDi Research America
Mountain View, CA, USA
yinhongchang@didichuxing.com

ABSTRACT

Digital threats such as backdoors, trojans, info-stealers and bots can be especially damaging nowadays as they actively steal information or allow remote control for nefarious purposes. A common attribute amongst such malware is the need for network communication and many of them use *domain generation algorithms* (DGAs) to pseudo-randomly generate numerous domains to communicate with each other to avoid being take-down by blacklisting method. DGAs are constantly evolving and these generated domains are mixed with benign queries in network communication traffic each day, which raises a high demand for an efficient real-time DGA classifier on domains in DNS log. Previous works either rely on group contextual/statistical features or extra host-based information and thus need long time window, or depend on lexical features extracted from domain strings to build real-time classifiers, or directly build an end-to-end deep neural network to make prediction from domain strings. Pros and cons exist for either way in experiments. In this paper, we propose several new real-time detection models and frameworks which utilize meta-data generated from domains and combine the advantages of a deep neural network model and a lexical features based model using the ensemble technique. Our proposed model obtains performance higher than all state-of-art methods so far to the best knowledge of the authors, with both precision and recall at 99.8% on a widely used public dataset.

CCS CONCEPTS

• **Security and privacy → Intrusion/anomaly detection and malware mitigation**; • **Computing methodologies → Neural networks**; **Ensemble methods**; *Supervised learning by classification.*

KEYWORDS

URL classification; DGA; Deep neural network; Recurrent neural network; Ensemble; Lexical features

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3316498

ACM Reference Format:
Shuaiji Li, Tao Huang, Zhiwei Qin, Fanfang Zhang, and Yinhong Chang. 2019. Domain Generation Algorithms detection through deep neural network and ensemble. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308558.3316498

1 INTRODUCTION AND REVIEW

With the rapid advancement in technology and the proliferation of digital devices in today's world, the amount of data being generated, shared and stored is more than ever before. Whilst this brings increased convenience in the form of, for example, online shopping and banking services, cybercriminals are also aware of the value of such data. Digital threats such as backdoors, trojans, info-stealers and bots are constantly evolving and can be especially damaging as they actively steal personal information or allow malicious attacks. One common attribute amongst such malware is the need for network communication. Such malware typically communicate with an attacker controlled *Command and Control* (C&C) server via C&C channels. In order for the malware to communicate with its C&C server, it has to know the server's IP address. One straightforward way is hard-coding a list of IP addresses or domain names in the malware. However, this renders the malware vulnerable to simple IP or domain *blacklisting*. In order to avoid detection and prevent being taken down, some malware families adopted *domain generation algorithms* (DGAs). DGAs are used to generate a large number of pseudo-random domains from a specified seed. The malware and C&C servers share the same seed and pseudo-random algorithm. For example, a DGA could use a date as its seed, as discovered by researchers who reverse-engineered the Torpig malware [27], or even a character of a tweet, as they later found in an upgraded version of Torpig. Malware employing DGAs can generate a set of new domains from each seed and attempt to connect to these domains. As the malware will continually try resolving and connecting to each domain, only a small subset of these domains need to be pre-registered by an attacker for this whole scheme to work. Because any seed value and algorithm can be used in DGA schemes, it is difficult to predict the domains that each malware will use. Therefore, this method effectively renders IP/domain blacklisting useless.

The domains being queried in a network can be acquired via logs of *Domain Name Service* (DNS) servers or a monitoring server. One can build a system to analyze each queried domain and give a

verdict, either malicious or benign, that guides further actions such as communication block or remediation. Due to the high volume of DNS traffic, and the target to block malware's C&C channels as early as possible, a *real-time* DGA classifier becomes important in such case, as blacklisting becomes inefficient, and reverse engineering [27, 28] [1] of the DGA might only work usefully for a while as attackers would update the seeds and algorithms. To build such classifiers, many features can be extracted, such as lexical features, contextual features, or host-based features [25]. Mixed feature type models exist in previous works [3, 32, 33]. However, as to a real-time system in DGA classification or general malicious URL detection, the most straightforward and commonly used features are still lexical features that describe the URL itself [1, 2, 5, 22, 24].

Artificial neural networks (ANNs) are a family of machine learning models that prove to achieve high performance on many difficult problems [12, 17, 21]. ANNs can be used to learn the patterns in the malicious URL strings and lexical features extracted. For example, Le et al. [23] used embedding to transform URLs to matrices and use a deep convolutional neural network (CNN) to classify malicious URLs. In a wide range of tasks involving time series and sequential data, *Recurrent Neural Networks* (RNNs) prove to be useful [15, 29]. RNNs are a class of neural networks with loops (so called *recurrences*) that recursively combine the input vectors with the network state vectors to produce a new state vector, thus allowing temporal information to persist throughout the learning process, and therefore successfully learn some patterns in sequential data. In particular, *long-short term memory* (LSTM) [14, 18] is a special kind of RNN frequently used for data with long-term temporal dependencies. It solves the *vanishing gradient* problem that a vanilla RNN might have [2] by employing cells that preserve information and use 3 "gates" (input, output and forget) to control the information flow. A trained RNN-based model like LSTM is fast in inferencing and prediction, and hence naturally suitable in real-time malicious URL or DGA classification. Bahnsen et al. [4] used LSTM networks directly to the URL string sequence to predict if the URL is a Phishing URL. In DGA classfication, since URLs are randomly generated, RNN models can be developed to learn and detect the "randomness" patterns. Woodbridge et al. [31] used a 128-units single layer LSTM (with input embedding layer) model to learn the "randomness" of different generation algorithms (such as *ramnit, shifu, suppobox, beecone, simba* etc.) and make binary predictions (DGA or non-DGA) or multi-class prediction (the specific algorithm prediction). They reached 98% true positive rate (TPR) and 0.1% false positive rate (FPR) on a public database of DGA from Bambenek Consulting [10]. Within an industry setting, assuming 1 million URLs are queried in the network on a daily basis, and 0.1% of them are real malicious DGAs. A 0.1% FPR will produce roughly 1,000 false positive cases, which might require further inspections from domain experts. We aim to reduce this workload, and at the same time, maintain the generalizability of the algorithm.

The main drawback of these two models is that they solely rely on an LSTM on the string itself, and neglect a wide range of lexical features that can be extracted and assist the classification, like n-grams, length, special characters, etc, as in [5, 13, 20, 22, 24, 26, 32, 34]. Potential failures caused by such drawback lie in two folds. First, sole URL string based deep learning model might "over-learn" the negatives and positives in training set, and generalize not so well in reality [3]. Furthermore, low interpretability of neural networks make it hard to improve or amend such models. Second, some lexical features, like n-grams, special characters, domain length, etc. might convey patterns that can really help the classification, yet these models did not embed them explicitly. RNN model might finally learn these helpful features in an implicit way with abundant data, but combining these features directly in the model might be more efficient. Also, combining string based RNN model and lexical features model might help to "neutralize" the flaws of each individual model and make performance improvement, as we will show in this work. In this paper, we propose a new framework that utilize RNNs and lexical features to form a mixed model to improve the DGA prediction performance. The rest of the paper is arranged as following: in Section 2, we will propose two RNN based models, probability network and representation network, an n-gram based model, and an ensemble model that combines these models and extra lexical features extracted. In Section 3, we will show our experiment settings. Experiment results and discussions will be in Section 4.

2 METHODOLOGY

Our methodology consists of two new deep neural network models and an n-gram-based Xgboost model. Each of them has its unique advantages and disadvantages, and they are finally integrated into an ensemble model that combines the best of all. We will demonstrate the prediction performance in Section 4.

2.1 RNN-based model

We propose two network architectures that are similar but motivated by two different concepts and thus learned differently — a probability network and a representation network.

2.1.1 Probability network. Intuitively, the whole deep learning model tries to capture the pattern in the URLs in order to make prediction of the probability that a given URL is DGA. A graphical depiction of our probability network model is shown in Fig 1. We combine extra lexical features like top-level-domain (TLD) and string length, two commonly used lexical features in previous works in malicious URL identification [3, 13, 20]. A TLD is one of the domains at the highest level in the hierarchical Domain Name System of the Internet. For example, in 'photos.google.com', 'com' is the TLD. String type inputs need to be tokenized first, and get truncated or padded to a uniform length. Subsequently, in the arena of machine learning, one-hot encoding and embedding are two techniques popular in transforming the tokenized string input to a reasonable numerical representation that can be fed into learning models. The one-hot encoding transforms all the character-wise tokens from the original input to sequences of binary representation. As a consequence, it can be slow and space-consuming with high redundancy. On the other hand, embedding condenses the information from the original sequence and can be optimized together with the subsequent models, but it can also lose some useful information

[1] After reverse-engineer, defenders can register anticipated domains and even hijack the C&C sever the attackers use, known as *sinkholing*

[2] the problem which make "memorizing" long-range dependencies difficult

[3] which indeed happened when we deploy such model, as false positives happen on benign domains rarely visited

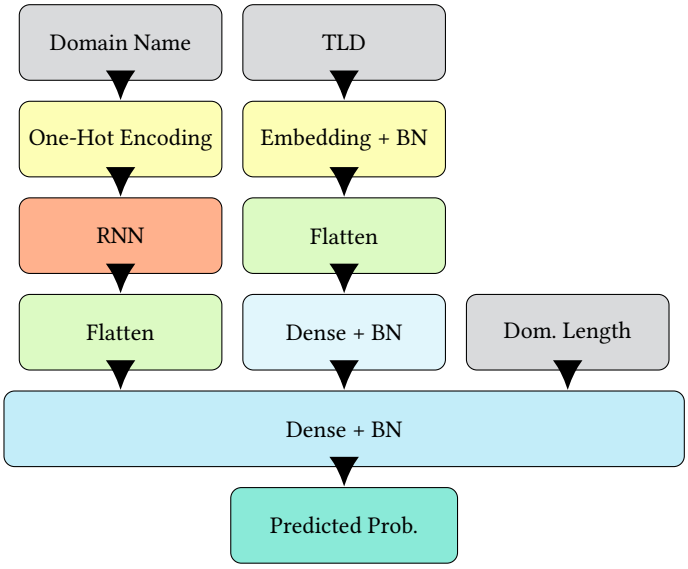


Figure 1: Proposed probability network based on RNN. "BN" here stands for the batch normalization.

from the original sequence. In our probability network, we feed in the full information from the domain name by using one-hot encoding. Meanwhile, we use an embedding layer to the TLD input, as a TLD can only take its value in a predefined list with thousands of legal TLDs [4] and thus is relatively less complex compared to domain names.

An RNN such as LSTM is later used to learn the DGA patterns in the domain name sequence. Its outputs, concatenated with the domain length and the output from TLD processing, are fed into one or multiple fully connected layers before the probability prediction is generated from a final *sigmoid* activation.

2.1.2 Representation network. For text type data problems, such as NLP tasks, pre-training a representation network using RNN models was proved to be effective for improving performances as downstream models can be built up based on these represented vectors [11]. Similarly, when we leverage an ensemble model as shown in Fig 2, we could build an RNN model that maps the URL string (and TLD) to a low-dimensional space, where DGAs are "clustered" and separated from non-DGAs, to ease the learning of the downstream ensemble model. In fact, our *probability network* can be seen as a 1-dimensional representation of the URL strings.

The representation network is constructed similarly to the probability network we proposed, but we intended to make it light-weight, which is very important when we have enormous training data. Similar to probability network, domain names and TLDs are separated and processed, and finally concatenated before feeding into fully connected layers. To make it light-weight, first we replace the one-hot encoder of domain names with an embedding layer to accelerate the rest of domain name learning; second we choose the GRU [9] to learn embedded domain names instead of an LSTM model[5]. Domain length input is canceled in this network. The final low-dimensional output vector of this network is saved as a group of new features for our ensemble model.

[4]Internet Corporation for Assigned Names and Numbers (ICANN) manages most of the legal TLDs

2.2 n-gram based model

In addition to the deep neural network models, we build an Xgboost [8] model to learn the pattern of DGAs. Similar to the idea of [33], we first construct a bigram occurrence count table for each possible character bigram from the training data. The intuition is that generative algorithms such as DGAs are unlikely to generate domain names that preserve a bigram distribution that is exactly the same to legitimate domain names. For a benign domain, it is common to see some "regular" combinations, such as 'go' within 'Google', or 'be' within 'YouTube', either of which contains legitimate lexical items or follow general pronunciation habit in English or some other languages. A malicious DGA, in contrast, tends to disrupt the order of characters[3] so as to circumvent the patterns captured in blacklists. Hence, a DGA domain with counts on less frequent character bigrams is likely to be detected through an n-gram based model.

Once we obtain the character bigram features, we can train a gradient boosting based classifier (Xgboost in this case) to make a prediction. The reason of choosing Xgboost over conventional gradient boosting decision-tree (GBDT), or Adaboost, is that Xgboost can be trained in distributed fashion using multi-cores, which drastically speeds up the training process. In [1], trigram features as opposed to bigram features were used to train a random forest (RF) DGA classifier. In our experiment, we found that bigram features outperform unigram or trigram features in all the evaluation metrics. We also show that gradient boosting based models perform better than RF regardless the choice of n in character n-grams ($n < 4$). Although both RF and Xgboost belong to ensemble learning methods, they differ in multiple aspects. For RF, the order of trees is not important since each tree is built independently using bootstrap samples of the training data. Xgboost, on the other hand, adds new trees to already built ones by continuously minimizing the pseudo-residual over decision-tree space. According to [6, 7], as long as the hyper-parameters are carefully tuned, gradient boosting based models are more robust than RF over a regular feature space with a dimension less than 4,000 because of the above difference in the training objective. In our case, the experiment results in roughly a thousand of bigram features.

2.3 Ensemble

As explained previously, RNN-based models are good at learning temporal information in sequential data. Our RNN-based models are featureless, with syntactic and semantic representations only taken into account implicitly through the training process. Global patterns in domain names and long term correlations among characters can be captured for malicious and benign domains. However, as mentioned in [31], the global "randomness" of the generated domains is hardly captured by the RNN if a training set has inadequate instances of some generation algorithms. Meanwhile, n-gram based model is better in detecting malicious domains that share character n-gram (local patterns, as opposed to full string patterns)

[5]Compared to LSTM, GRU has no memory unit and is therefore faster to train.

distributions with known DGAs in the training set. It does not retain any semantic meaning of a sequence. Therefore, it is possible that a benign domain is classified as DGA by the n-gram model because it shares some bigram tokens with some malicious domains in the training set even though the bigram's position in the string, or leading/following bigrams are different. This yields a high false positive rate (FPR), which costs more further investigations and, of course, higher labor costs, as we found in production test.

To mitigate the influence of global and local information, and to maintain the efficacy of the model with lower false positive rate (FPR) and false negative rate (FNR), we propose a novel ensemble model that encompasses the virtue of both RNN-based and n-gram based models with some extra URL-based features. Inspired by [3–5, 26, 32, 33] etc., the extra features we make from a domain include:

- Length of name s ($s \in$ {domain name, TLD})
- Whether the length of a name s is larger than a threshold c_s ($s \in$ {domain name, TLD})
- Number of numerical characters contained in domain name, tld (also whether tld contains any numerical character)
- Number of special characters contained in domain name and whether TLD contains any special character

A graphical description of the ensemble model is shown in Fig 2. We concatenate the DGA probability predictions from RNN-based model and the probability predictions from Xgboost model described above, together with the additional features we make out from domain strings, and then feed them to a top logistic regression model for the final DGA probability prediction

$$\begin{aligned} p_i &= \frac{1}{1+\exp\left[-(\mathbf{w}^\top \mathbf{x}_i + b)\right]} \\ &= \frac{1}{1+\exp\left[-(w_1 x_{1i} + w_2 x_{2i} + \ldots + w_d x_{di} + b)\right]} \end{aligned} \tag{1}$$

where $\mathbf{w}$ is the linear weights, $\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{di})^\top$ is the concatenated vector containing extra features and output from the upstream components, and b is the model bias.

At the training stage, both RNN-based model and n-gram based model are trained separately first. On the second stage, their predictions on training data are used in the top logistic regression model training to solve for the ensemble coefficients. By default the coefficients are regularized with ℓ_2 norm, and the coefficient of the regularization term is found by grid search.

3 EXPERIMENT SETUP

In this section, we describe the implementation specifics of all the models developed in Section 2.

3.1 RNN-based model

Probability network. In the probability network model, as an experiment, we implement an LSTM network which is composed of one layer of 256 units LSTM. LSTM output sequences are flattened before being sent to the fully connected dense layers. TLDs are first embedded to 5-dimensional vectors and then fed to a 32 units dense layer with rectifier activation. LSTM model output, TLD output and domain length were concatenated and then fed to a 32 units dense layer with rectifier activation. The output layer is a 1 unit layer with sigmoid activation. Batch normalization layers are added to some layers as shown in Fig 1. We maximize the likelihood of the training instances by optimizing the binary cross-entropy loss.

Representation network. In the GRU based representation model, domain names are first embedded to 63-dimensional vectors and fed into an RNN that is composed of one layer of 256 GRU hidden units. TLDs are embedded to 7-dimensional vectors and fed to a 32 units fully connected layer. GRU output matrices are then flattened and concatenated with the dense representations of TLDs and fed together into a 64 units dense layer. Two subsequent dense layers, with 32 and 1 dense units each, are employed to transform joint dense representations into probability measures. Rectifier activation is used on all intermediate dense layers, and the last layer is activated through a sigmoid function. Once the representation network is trained, output vectors of the second to last layer (32-dimensional) is exported for the ensemble model.

3.2 n-gram based model

For n-gram based models, we test the performance of RF and Xgboost classifiers on unigram, bigram, and trigram features of domain names, respectively. Character n-gram features are extracted on training set, with random holdout test set to measure the generalizability to detect DGA from benign URLs. To reduce potential over-fitting, for each classifier we grid search the hyper-parameters including number of trees and maximum depth of each tree. We choose logistic loss as the objective function of Xgboost and fix the learning rate to 0.1. Information gain entropy is used as the node splitting criterion for random forest.

3.3 Ensemble layer

Once the RNN-based model and n-gram based model are thoroughly trained, we build a top-level logistic regression model taking input from RNN-based model output and n-gram based model output, together with some URL-based features, as shown in Fig 2. The top-level feature space contains either 10 features.The threshold $c_s = \{45$ (domain names), 7 (TLDs)$\}$, and special characters include multiple symbols such as underscore and period. The layer weights are regularized with ℓ_2 norm and the penalty parameter is grid-searched through cross-validation. From now on, we call the generated features, corresponding to the definition in Section 2.3, as `domain_len, tld_len, is_dom_ge_45, is_tld_ge_7, #_num_char, #_num_char_in_tld, is_num_in_tld, #_spec_char_in_tld, is_spec_in_tld` for simplicity.

4 RESULTS

4.1 Overall performance

Table 1 concludes the test performance of all the models we tested in this paper. Previous popular used SVM-based models presented in [16, 30] did not even generalize well (precision and recall lower than 97%) on smaller test sets (less than 2,000 malicious DGAs), and were therefore excluded from our baseline experiments. For the LSTM model proposed in [31], we used the exactly same setting and reached a similar result as declared in the paper. The dataset we used was composed of benign URLs (from Alexa top 1M domains) and DGAs (from Bambenek Consulting[10], about 942k). 70% of

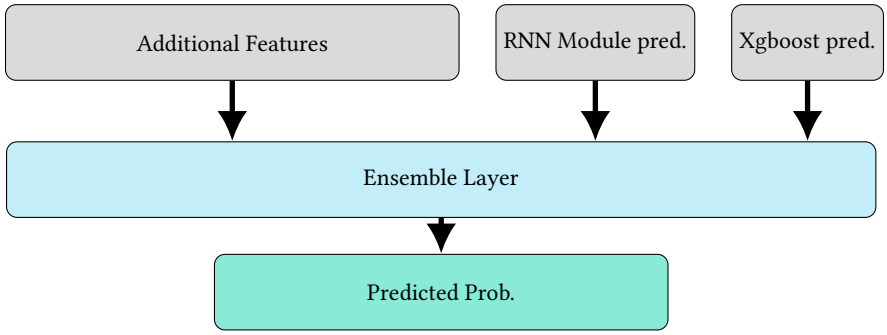


Figure 2: Ensemble model

Table 1: Experiment Results

		Evaluation Metrics				
	Model	Precision	Recall	F1-score	Accuracy	AUC
RNN based	LSTM proposed in [31]	0.9826	0.9819	0.9822	0.9822	0.9970
	Probability Network	0.9940	0.9930	0.9935	0.9937	0.9995
N-gram based	RF (unigram)	0.9637	0.9624	0.9628	0.9629	0.9919
	RF (bigram)	0.9737	0.9772	0.9727	0.9728	0.9965
	RF (trigram)	0.9482	0.9424	0.9434	0.9438	0.9928
	Xgboost (unigram)	0.9669	0.9661	0.9664	0.9665	0.9934
	Xgboost (bigram)	0.9854	0.9850	0.9852	0.9852	0.9986
	Xgboost (trigram)	0.9618	0.9611	0.9614	0.9614	0.9914
Ensemble	Ensemble-probability	**0.9982**	**0.9982**	**0.9982**	**0.9982**	**0.9999**
Ensemble	Ensemble-represent.	0.9967	0.9967	0.9967	0.9967	0.9998

the dataset was used as the training set, and the rest used as the testing set. Our probability network outperforms the LSTM network proposed in [31], with 99.40% precision and 99.30% recall. Among all the n-gram based models, our Xgboost classifier outperforms the RF classifiers proposed by Ahluwalia et al. [1] on all three cases ($n = 1, 2, 3$), with $n = 2$ the best configuration. Overall, the n-gram, tree-based models underperform the probability network we proposed. The ensemble-probability model outperforms the rest models, and show significant improvements over methods in studied literature. The ensemble-probability model achieves 99.82% precision and recall. To the authors' best knowledge, our model obtains the best performance on DGA classification.

Besides, based on Table 1, the FPR of our ensemble-probability model is 0.17%, almost ten times smaller than the LSTM network(1.60%) proposed in [31] [6]. The level of potential human involvement (i.e. domain expert inspections) is effectively lessened. We have also tested the ensemble model on various test sets collected from different time periods (March - May 2018), and it achieved similar performance, a sign of well-preserved generalization ability.

4.2 Case studies

[6]Note here the FPR we get is different from what [31] claimed in their paper; we cannot reproduce the 0.1% FPR in our experiment using our dataset and the code snippet provided in [31]

4.2.1 LSTM-based probability networks. LSTM takes its input from the domain character sequence, and returns an output vector, or a sequence of cell output vectors if we choose to. We wonder: 1) how the prediction is formed as the characters were fed into the LSTM sequentially, and 2) in which point the model raises or decreases its prediction. For the first question, Karpathy et al. [19] visualized LSTM cell outputs in heatmaps and found units (or dimensions, channels) of cell states that have interpretable use in prediction. We could do so as well to explore our probability network, but it turns out there seems no specific human interpretable meaning of any unit that can apply to all domains. For the second question, we can expand subsequences of a domain character-by-character from the first character to the whole string, feed these subsequences to the LSTM based model and observe how the predicted probability is changing. This experiment was conducted on our probability network and the LSTM in [31] respectively on several selected domains and Fig 3 shows the result. The probability network seems to be able find some patterns that are meaningful in human language and thus clear the DGA suspicion of benign domains while the LSTM in [31] fails to do so for these cases. For example, in 'jpopsuki.eu' case, the probability network predicts 'jpopsuk.eu' as DGA, but when the last 'i' completes a meaningful subsequence 'suki'[7] the probability network drops its probability prediction to 'benign'. Similar phenomena are observed when 'n' completes 'xueyuan', 'ng' completes 'xing'[8] and 'o' completes 'pro'. The exact opposite things happen for DGAs as well — when 'n' makes 'nation' and 'ter' makes 'master', the probability network identifies the word and changes its prediction to DGA, probably because in training set there are similar DGAs composed by these words.

However, in next discussion, we will show that for patterns not in the training set, like some Chinese Pinyin sequence, the probability network might make mistakes and in such cases ensemble with lexical feature models can help to make the correct predictions.

4.2.2 Probability Network vs. Ensemble. As shown in Table 2, the weights w_1 (corresponds to the output probability from the probability network) and w_2 (corresponds to the output probability of the n-gram model) are 7.3741 and 9.0226 respectively. The n-gram Xgboost model exhibits a relatively larger impact than the probability network on final predictions of the ensemble layer, even though the

[7]Here 'suki' means 'like/love' in Japanese
[8]Both 'xueyuan' and 'xing' are meaningful in Chinese Pinyin

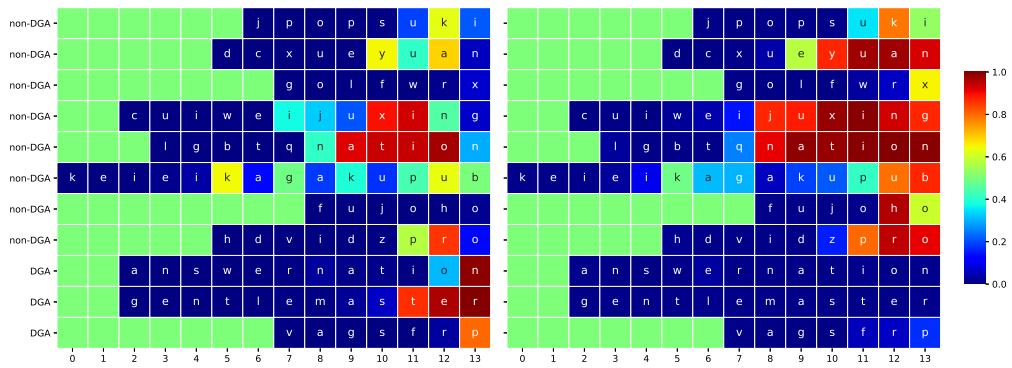


Figure 3: Comparison of LSTM based models on selective domains. The y-axis shows the true label of each domain. The left heatmap shows the prediction of our probability network when the domain is feed in as subsequences expanding character-by-character from the first character to the entire string, and the right graph shows the prediction of the model in [31]

Table 2: Weights of the ensemble layer in the ensemble-probability model

	Prob. Network	N-gram	domain_len	tld_len
Weight	7.3741 is_dom_ge_45	9.0226 is_tld_ge_7	0.0343 #_num_char_in_tld	-0.1086 #_num_char
Weight	-0.00085 is_num_in_tld	0.0037 is_spec_in_tld	-0.0043 #_spec_char_in_tld	-0.1214 b
Weight	-0.0036	-0.9690	-0.9648	-7.4072

probability network outperforms the n-gram model when tested separately.

In the test set, we examined selective URLs, on which the probability network and the n-gram based Xgboost model of the ensemble-probability model predicted differently. Contribution of each component to the final prediction is shown in Table 3. The first two columns show the probability network and n-gram model output probabilities respectively. The third and fourth columns of Table 3 show the component values of these two models (i.e., w_1x_{1i} and w_2x_{2i} respectively), while the fifth column is the sum $w_1x_{1i} + w_2x_{2i}$. The last column shows whether our prediction is correct or wrong. Clearly, a larger $w_1x_{1i} + w_2x_{2i}$ component value means a higher final predicted probability produced by the ensemble model, indicating that the two probability features dominate the top layer's feature space in terms of feature importance.

Diving into the sample cases, we saw that RNN model could correctly identify some benign URLs composed of letters that look "random" and bear no apparent semantic meanings, such as abbreviations from different languages mixed, like 'ccpitnmg' in the sample cases. These domain names look like real DGAs and contain "uncommon" n-grams within the benign cases of the training set from a character perspective. Oppositely, when DGA looks "meaningful" with "common" n-grams that can be found in English (or directly a combination of meaningful words), like 'meskia', 'gentleshould' or 'basketmillion', N-gram model might fail easily. RNN model can, in this case, learns the hidden representation of the domains as a whole through LSTM. N-gram model fails to detect this underlying distribution since bi-gram features are disjoint — the encoded token values are primarily used for nodes split, with weak correlations among n-grams through learning.

However, we noticed that RNN model could have trouble in decoding some sequence patterns, for example, some of the Chinese Pinyin based domains, such as 'xizanglvyou', 'jiafalvshi' and 'ixueyi' [9]. It tends to output a high probability, which make us believe that the LSTM does not generalize its good performance on other similar semantic based domains (like 'gentleshould') to these domains. A probable reason is that our training set of benign URLs (Alex top 1M) contains inadequate examples for Chinese Pinyin patterns. At each cell of the LSTM, the emitted hidden vector contains cell values of current input character with information of previous characters[14, 18]. To always obtain a representative vector that captures the underlying sequence pattern over hidden units, LSTM layers should be trained on a data set that contains enough input sequences with diversified distributions. The authors in [31] have also claimed that lack of training instances generated by some rarely used generation algorithms lead to pool generalization of its LSTM-based model in real production system. For these cases, our N-gram based model makes the right predictions based on their character bi-gram counts.

These cases show the improved robustness of our ensemble model, as compared to individual RNN based and n-gram based model. On some patterns and cases, the two component models can be complementary. Moreover, under this framework, we can easily explore more (complementary) component models and add them to the ensemble.

4.3 Representation of DGAs

Fig 4 shows a 2-dimensional visualization of representation network output for selected test data using t-SNE. Overall, the representation network successfully created the low-dimensional representations of the test domains that are easy to separate and classify. Most benign and malicious cases clearly form its own cluster except for some cases in the lower portion of the graph. This explains why

[9] literally translate: 'Tibet tourism', 'family law lawyer' and 'I learn skill'. The 'v' in 'lv' stands for 'ü' when typewriting Pinyin using standard keyboard

Table 3: Selected ensemble-probability model's prediction components of sample URLs

Domain	RNN prob.	N-gram prob.	RNN comp.value	N-gram comp.value	RNN+N-gram comp.value	Ensemble prob.	True label	Pred.*
xizanglvyou.org	0.9998	0.0134	7.3725	0.1205	7.4930	0.5343	0	✗
jiafalvshi.com	0.8061	0.0952	5.9445	0.8590	6.8036	0.3575	0	✓
ixueyi.com	0.7309	0.0914	5.3897	0.8248	6.2145	0.2121	0	✓
ccpitnmg.org	0.2480	0.6315	1.8291	5.6973	7.5264	0.5170	0	✗
meskia.biz	0.9823	0.0047	7.2434	0.0425	7.2859	0.4400	1	✗
ngxdlr.org	0.2158	0.9670	1.5916	8.7247	10.3164	0.9421	1	✓
ysmyhuh.net	0.9659	0.1778	7.1224	1.6044	8.7268	0.7745	1	✓
gentleshould.net	0.9982	0.0124	7.3612	0.1117	7.4729	0.5378	1	✓
basketmillion.net	0.9948	0.1011	7.3358	0.9123	8.2481	0.7234	1	✓

*: ✓ means our prediction is correct and ✗ means our prediction is wrong.

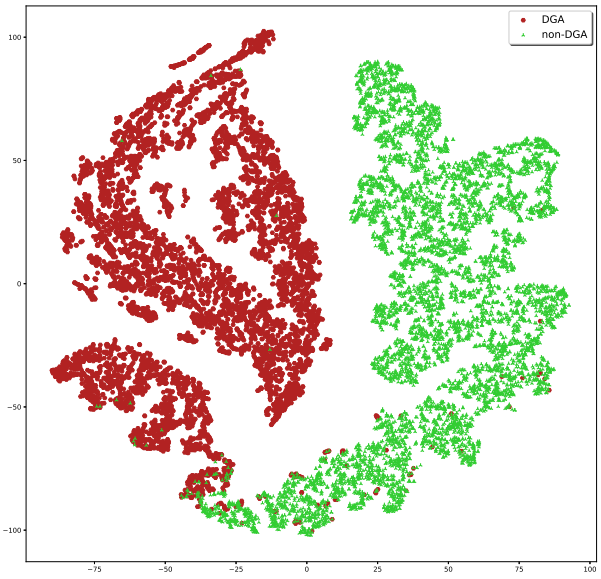


Figure 4: A t-SNE embedding over representation vectors output from GRU based *representation network*. For each group (DGA vs. non-DGA), we randomly selected 8,000 domains from the test data of the group.

ensemble-representation model obtains a very promising performance close to ensemble-probability. Bear in mind that the representation network is initially designed as a much lighter-weight model compared to the probability network, and indeed, the running time in our experiments of the representation network training is only 1/10 of the running time of the probability network training using same machine. For situations where computation power or time frame is limited, we suggest using ensemble-representation model for better efficiency-performance trade-off. Besides, this representation network can be easily plugged-in other learning models and frameworks as part of input processor.

Table 4: Selected edge cases where extra features are influential on final prediction

domain	RNN prob	N-gram prob	RNN+ N-gram comp_val	All comp _val*	Ensem. prob	True Label
eyeota.net	0.9387	0.0614	7.4761	7.3562	0.4872	0
vorobatorakhtopl .com	0.0014	0.8136	7.3511	7.5742	0.5417	1
houmemapoig gofjogra.com	0.0000	0.8146	7.3502	7.6762	0.5669	1

*: As the intercept is -7.4072 from Table 2, if an All comp_val is larger than 7.4072, the model will give a positive prediction and vice versa.

4.4 Extra Features

From Table 2, most extra features have relatively small coefficients after training, and indeed for most of test cases, the predicted label is mostly determined by RNN-based model output(s) and n-gram based model output. However, in some "edge" cases where RNN-based model and n-gram based model disagree and almost "neutralize" each other, the extra lexical features take effect. Table 4 shows 3 cases where such disagreement happens and the extra features help make the correct predictions (mainly length related extra features are making difference in these cases).

5 CONCLUSION

In this paper, we built RNN-based deep neural network models for DGA prediction. Together with n-gram based machine learning model for DGA classification and extra lexical features we made, we proposed to build a top ensemble layer that combines the advantages of the n-gram based model, assistive lexical features and the full-string based model (here the deep neural network models) while at the same time alleviating the disadvantage of each individual component when it is used separately to a great extent. As a result, first, the LSTM based probability network outperforms the literature [31] LSTM model by almost 1 percent in both precision and recall; second, without loss of the generalization capability, the

ensemble model obtains a new state-of-the-art result of 99.82% precision and recall, higher than the standalone n-gram based model and RNN-based probability network model we proposed.

In the discussion of the result, we show that either the full-string based deep learning model or the n-gram based model might not be able to capture all DGAs, as DGA is versatile and the task is complex. We suggest using an ensemble model to balance the two types of models. Our work shows the virtue of each part and proposes a viewpoint of combining them together in a good way. We believe this thought is more important than the specific ensemble implementation (i.e. logistic regression) we propose here, as it is just one way of implementing the thought, and we expect more future works keep revealing the pros and cons of different types of models in DGA prediction and the approaches of composing them to a better meta-model. Our representation network proves to be effective as well as it separates DGAs and non-DGAs in representation space. Meanwhile, DGA is also keep evolving, for example, we find DGAs based on random word sequences instead of character sequences. This brings need for future works on word/character mixed DGA prediction models.

REFERENCES

[1] Aashna Ahluwalia, Issa Traore, Karim Ganame, and Nainesh Agarwal. 2017. Detecting Broad Length Algorithmically Generated Domains. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Springer International Publishing, 19–34.

[2] Iraj Sadegh Amiri, Oluwatobi Ayodeji Akanbi, and Elahe Fazeldehkordi. 2014. *A machine-Learning approach to phishing detection and defense.* Syngress.

[3] Manos Antonakakis, Roberto Perdisci, Yacin Nadji, Nikolaos Vasiloglou, Saeed Abu-Nimeh, Wenke Lee, and David Dagon. 2012. From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware.. In *USENIX security symposium*, Vol. 12.

[4] Alejandro Correa Bahnsen, Eduardo Contreras Bohorquez, Sergio Villegas, Javier Vargas, and Fabio A González. 2017. Classifying phishing URLs using recurrent neural networks. In *Electronic Crime Research (eCrime), 2017 APWG Symposium on*. IEEE, 1–8.

[5] Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. 2010. Lexical feature based phishing URL detection using online learning. In *Proceedings of the 3rd ACM workshop on Artificial intelligence and security*. ACM, 54–60.

[6] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. 2008. An empirical evaluation of supervised learning in high dimensions. In *In International Conference on Machine Learning (ICML*. 96–103.

[7] Rich Caruana and Alexandru Niculescu-Mizil. 2005. An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics. In *In Proc. 23 rd Intl. Conf. Machine learning (ICML'06)*. 161–168.

[8] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[10] DGA Domain Feed Bambenek Consulting. 2018. https://osint.bambenekconsulting.com/feeds/dga-feed.txt. Accessed: 2018-03-06.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 69–78.

[13] Sujata Garera, Niels Provos, Monica Chew, and Aviel D Rubin. 2007. A framework for detection and measurement of phishing attacks. In *Proceedings of the 2007 ACM workshop on Recurring malcode*. ACM, 1–8.

[14] Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* 12, 10 (Oct. 2000), 2451–2471. https://doi.org/10.1162/089976600300015015

[15] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 38 (March 2013), 6645–6649.

[16] F. Haddadi and A. N. Zincir-Heywood. 2013. Analyzing string format-based classifiers for botnet detection: GP and SVM. In *2013 IEEE Congress on Evolutionary Computation*. 2626–2633. https://doi.org/10.1109/CEC.2013.6557886

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[19] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078* (2015).

[20] Pranam Kolari, Tim Finin, and Anupam Joshi. 2006. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County. Also available as technical report TR-CS-05-13.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[22] Anh Le, Athina Markopoulou, and Michalis Faloutsos. 2010. PhishDef: URL Names Say It All. *CoRR* abs/1009.2275 (2010). arXiv:1009.2275 http://arxiv.org/abs/1009.2275

[23] Hung Le, Quang Pham, Doyen Sahoo, and Steven CH Hoi. 2018. URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection. *arXiv preprint arXiv:1802.03162* (2018).

[24] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2009. Identifying Suspicious URLs: An Application of Large-scale Online Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. ACM, New York, NY, USA, 681–688. https://doi.org/10.1145/1553374.1553462

[25] Doyen Sahoo, Chenghao Liu, and Steven CH Hoi. 2017. Malicious URL detection using machine learning: A survey. *arXiv preprint arXiv:1701.07179* (2017).

[26] Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, and Stefano Zanero. 2014. Phoenix: DGA-based botnet tracking and intelligence. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 192–211.

[27] Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. 2009. Your Botnet is My Botnet: Analysis of a Botnet Takeover. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09)*. ACM, New York, NY, USA, 635–647. https://doi.org/10.1145/1653662.1653738

[28] Brett Stone-Gross, Marco Cova, Bob Gilbert, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. 2011. Analysis of a botnet takeover. *IEEE Security & Privacy* 9, 1 (2011), 64–72.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173

[30] Zhen Wang, Zhongtian Jia, and Bo Zhang. 2018. A Detection Scheme for DGA Domain Names Based on SVM. (2018). https://doi.org/10.2991/mmsa-18.2018.58.

[31] Jonathan Woodbridge, Hyrum S Anderson, Anjum Ahuja, and Daniel Grant. 2016. Predicting domain generation algorithms with long short-term memory networks. *arXiv preprint arXiv:1611.00791* (2016).

[32] Sandeep Yadav, Ashwath Kumar Krishna Reddy, AL Reddy, and Supranamaya Ranjan. 2010. Detecting algorithmically generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 48–61.

[33] Sandeep Yadav, Ashwath Kumar Krishna Reddy, AL Narasimha Reddy, and Supranamaya Ranjan. 2012. Detecting algorithmically generated domain-flux attacks with DNS traffic analysis. *IEEE/Acm Transactions on Networking* 20, 5 (2012), 1663–1677.

[34] Wen Zhang, Yu-Xin Ding, Yan Tang, and Bin Zhao. 2011. Malicious web page detection based on on-line learning algorithm. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, Vol. 4. IEEE, 1914–1919.

Hi Doppelgänger: Towards Detecting Manipulation in News Comments

Jan Pennekamp
Communication and Distributed Systems
RWTH Aachen University, Germany
jan.pennekamp@comsys.rwth-aachen.de

Martin Henze
Communication and Distributed Systems
RWTH Aachen University, Germany
martin.henze@comsys.rwth-aachen.de

Oliver Hohlfeld
Communication and Distributed Systems
RWTH Aachen University, Germany
oliver.hohlfeld@comsys.rwth-aachen.de

Andriy Panchenko
IT Security
Brandenburg University of Technology, Germany
andriy.panchenko@b-tu.de

ABSTRACT

Public opinion manipulation is a serious threat to society, potentially influencing elections and the political situation even in established democracies. The prevalence of online media and the opportunity for users to express opinions in comments magnifies the problem. Governments, organizations, and companies can exploit this situation for biasing opinions. Typically, they deploy a large number of pseudonyms to create an impression of a crowd that supports specific opinions. Side channel information (such as IP addresses or identities of browsers) often allows a reliable detection of pseudonyms managed by a single person. However, while spoofing and anonymizing data that links these accounts is simple, a linking without is very challenging.

In this paper, we evaluate whether stylometric features allow a detection of such doppelgängers within comment sections on news articles. To this end, we adapt a state-of-the-art doppelgänger detector to work on small texts (such as comments) and apply it on three popular news sites in two languages. Our results reveal that detecting potential doppelgängers based on linguistics is a promising approach even when no reliable side channel information is available. Preliminary results following an application in the wild shows indications for doppelgängers in real world data sets.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; • **Security and privacy**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

online manipulation; doppelgänger detection; stylometry

ACM Reference Format:
Jan Pennekamp, Martin Henze, Oliver Hohlfeld, and Andriy Panchenko. 2019. Hi Doppelgänger: Towards Detecting Manipulation in News Comments. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316496

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316496

1 INTRODUCTION

In the recent past, a new trend of opinion manipulation emerged on the Internet. As a consequence, media outlets increasingly report about fake news, propaganda, and trolls (online identities manipulating other users with their postings) [23]. While most of these concepts are well-known from a pre-Internet era, they have become more dangerous than before as a significant amount of users relies on the Internet as their primary information source [14]. Consequently, the validity of online information is at stake, especially since users are ubiquitously confronted with user-generated content, e.g., in forums, social networks, and comments to news articles. Governments, organizations, and companies might exploit this situation to manipulate users' opinions in their favor [34].

Traditionally, opinion manipulation on the Internet has spread through social media [4], e.g., during the Ukraine crisis [25] or the 2016 US presidential election [7]. For example, Facebook reported on closing multiple 10000 accounts in a single week due to suspicious activities [15]. Likewise, Twitter identified more than 50000 manipulating accounts during the presidential election [30]. Recently, we observed a shift of opinion manipulation towards the comment sections of news sites [21]. This is an especially alarming trend, since over 90 % of US Americans retrieve news online and 63 % of online news papers allow comments on articles [10]. For example, *the Guardian* reports that their moderators identify up to 250 opinion manipulating comments under a *single* article [12].

To counter the immediate threat to society introduced by online manipulation, related work usually relies on manual flagging by ordinary users (e.g., [20, 28]) or side channel information, such as IP addresses [19]. Clearly, the latter can be circumvented by applying anonymization techniques. In contrast, we focus on a scenario where manipulators actively *try* to conceal their identity, e.g., by hiding IP addresses or using fake email accounts.

In such a setting, one important building block to detect misinformation in the comment sections of news sites is the ability to detect *doppelgängers*, i.e., people using multiple identities with the goal to amplify the influence of the opinion they spread. As these trolls try to remain undiscovered, side channel information, such as IP addresses or email accounts, is often not reliable [17]. Thus, doppelgänger detection has to mainly rely on information which is difficult to spoof. While anonymizing or spoofing technical communication aspects (e.g., IPs) is feasible, consistently maintaining separate writing styles is hard [2]. In this regard, we explore the

applicability of characteristics of written texts to facilitate this task. Detecting doppelgängers based on written text has been studied before for longer texts, such as articles or blog posts [24]. However, comments on news sides are considerably shorter, rendering detection of doppelgängers much more challenging.

In this paper, we target the question on whether detecting doppelgängers in comment sections of news sites is feasible merely by leveraging linguistic features. Ultimately, this approach assists providers of news sites to detect attacks on the public opinion even if manipulators purposely disguise their behavior by obfuscating client-specific data, such as IP addresses, used browsers, or email addresses. More precisely, our contributions are as follows:

(1) We extend a state-of-the art doppelgänger detector [3] with additional features to account for comment-specific characteristics (short and informal text with only 50 to 1000 characters).
(2) We evaluate the feasibility of doppelgänger detection for short news comments solely based on linguistic features using a random subset of more than 4.8 million authentic news comments from three popular news sites in two languages. Our results show that our approach is able to detect the presence of doppelgängers with high detection accuracy, even when the number of doppelgängers is unknown. It also works well on two languages, German and English.
(3) We briefly hint at preliminary results of applying our approach to detect doppelgängers in the wild. Here, we find indications for user accounts constituting doppelgängers.

With our work, we contribute a building block to methods for fighting public opinion manipulation in online media.

2 RELATED WORK & BACKGROUND

To remedy the immediate threat resulting from opinion manipulation, especially in the context of online news comments, providing approaches that reliably *detect* opinion manipulation is imperative. Only if we are able to detect opinion manipulation, we can take appropriate countermeasures, such as flagging, moderating, or deleting comments. In the following, we first provide an overview of different approaches to detect opinion manipulation on the Internet. Subsequently, we take a closer look at approaches to detect doppelgängers, i.e., opinion manipulators who utilize more than one user account to amplify the credibility of their agenda, and especially those approaches that focus on stylometric features.

2.1 Detection of Opinion Manipulation

The first approaches tackling the detection of opinion manipulation were concerned with identifying spam in product reviews (e.g., [6, 19]). Subsequent research shifted towards social networks. E.g., Vosoughi et al. [31] conduct a longitudinal analysis of news story spreading on Twitter while Zannettou et al. [33] additionally include Reddit and 4chan into their analysis of manipulation.

On a related note, Zannettou et al. [32, 34, 35] investigate different aspects of the behavior of government-paid trolls. Kumar et al. [18] analyze the behavior of trolls in forums based on IP addresses and which discussions specific accounts participate in. Mihaylov et al. [20] expose paid trolls in comment sections based on their commenting behavior. Subsequently, they extend their approach [22] to also incorporate linguistic features to discover paid trolls. However, they only target spamming or annoying trolls. Consequentially, their approach would likely fail to discover user accounts with the focus of spreading misinformation.

2.2 Detection of Doppelgängers

To detect opinion manipulation, especially in the context of news sites, one important building block is the ability to detect doppelgängers, i.e., manipulators who operate multiple accounts to increase the impact of the opinion they express in their comments. Existing approaches to detect doppelgängers can be coarsely classified into approaches that either rely on metadata (e.g., time stamps or the topics that accounts are active in) or stylometric features (e.g., lexical, syntactic, and domain-specific writing style characteristics).

Using Metadata to Detect Doppelgängers. The most intuitive approach to detect doppelgängers relies on metadata such as time stamps of the published messages. Zheng et al. [36] detect doppelgängers in forums based on their activity frame and the topics they posted messages in. Johansson et al. [17] introduce the notion of "timeprints" to detect doppelgängers in social networks. Such a timeprint describes the account's activity pattern in a fixed set of features. The authors claim that the activity patterns are similar across accounts handled by the same manipulator.

Using Stylometry to Detect Doppelgängers. In contrast to approaches that rely on metadata, stylometric approaches focus solely on the written text and strive to identify text snippets that originate from the same author. This difference becomes especially relevant as manipulators try to avoid detection [17], hence rendering metadata, e.g., activity patterns, useless.

As a foundation to detect doppelgängers solely based on their written text, all approaches rely on so-called *stylometric features*, i.e., information that can (easily) be extracted from written text that, ideally, is unique for each individual author. Abbasi et al. [1] defined the most-relevant set of stylometric features, called *writeprints*, to model all aspects of writing style, especially targeting programming code snippets, ebay buyer/seller feedback, emails, and chat messages. These stylometric features cover lexical, syntactic, structural, content, and idiosyncratic aspects of written text. To also account for special characteristics of modern online communication, related work proposed to extend the writeprints set with additional domain-specific features, such as the sentiment of words [11], the number of received votes [20], and the usage of emoticons [11]. Authorship attribution is a well-established field of research and related work achieves promising results on longer texts, such as book chapters and blog posts [29]. Only recently, focus shifted towards shorter messages (as they are prevalent in news comments), e.g., in the context of Twitter [8, 26]. However, when considering news comments, existing approaches only rely on stylometric aspects to a very limited extent [22].

Using different sets of features, approaches from related work apply stylometry to detect doppelgängers in different settings and types of text. For example, Almishari et al. [5] demonstrate the linkability of Twitter accounts over time based on two simple features. They only consider unigrams and bigrams (n-grams measure the frequency of character sequences) to match doppelgängers. Afroz et al. [3] compile a subset of lexical and syntactic features from the

writeprints feature set and extend it with a feature expressing the leetspeak percentage to detect doppelgängers in blogs and forums. They input their features into the Doppelgänger Finder algorithm.

Doppelgänger Finder. This algorithm [3] is an approach to compare the writing styles of different authors with each other to obtain a similarity score for each pair of authors. The algorithm consists of three steps:

(1.) selecting meaningful features from the complete set of features by conducting a principal component analysis (PCA),
(2.) applying logistic regression to calculate the similarity of the writing style to all other authors for each author, and
(3.) combining these similarity scores to a single score for each pair of authors.

To interpret these scores, a threshold depending on various properties of the dataset [3] must be applied. Afroz et al. [3] recommend a manual analysis to identify such a threshold.

The approach taken by Doppelgänger Finder is especially relevant for our work, as the general algorithm is applicable to different means of communication. Furthermore, its implementation is freely available and thus an ideal foundation for our work. More specifically, in this work, we extend the overall process of comparing stylometric features using Doppelgänger Finder with additional features and an adaptable threshold metric to apply it for detecting doppelgängers in news comments, which often are shorter than postings in blogs and forums.

3 DETECTING DOPPELGÄNGERS IN NEWS COMMENTS

To detect doppelgängers in news comment sections, we present an approach which takes a set of user accounts, each having a pseudonym or user name, as input and generates a list of potential pairs of doppelgängers within this set of accounts as output. As such, we provide a check whether a single individual publishes texts with multiple accounts solely based on stylometric information. Our approach consists of three building blocks:

(1.) an augmentation of a state-of-the-art stylometric feature set [1] to adapt to modern online communication,
(2.) an application of a state-of-the-art doppelgänger detector [3] to retrieve similarity scores for each pair of accounts, and
(3.) an automated threshold metric to categorize the obtained similarity scores into potential doppelgängers and regular accounts.

We now present each of these building blocks in detail.

3.1 Stylometric Features for News Comments

As a foundation to detect doppelgängers based on written text, we first craft a stylometric feature set. We augment features proposed by different approaches from related work [11, 20] with own custom features specific to comments in news sections and create a unique feature set that – as the evaluation shows – is efficient in our targeted scenario.

Writeprint. We rely on the unmodified extended writeprints feature set consisting of 21 categories, such as word length distribution, part-of-speech tags, and word n-grams. Their default set of more than 27.6 k features (as we could not reproduce the 2300 part-of-speech tags, we follow the approach of Doppelgänger Finder [3] with 45 unigrams) is enlarged by the combination of occurring bigrams and trigrams. By default, we include all combination which accounts for more than 93 k additional features. While Doppelgänger Finder [3] uses only a reduced set of 8 categories ($\approx$ 1.1 k features), we implement the complete set.

Adaption for the Web. To adjust this feature set to the characteristics of modern online communication, we also include various extensions from related work [11, 20]. In particular, we decided to include an additional set of 165 features (f.) expressing:

(i) the sentiment of comments [11] (4 f.: positivity and sensitivity per word and sentence),
(ii) the activity periods of users [20] (4 f.: weekdays or weekends, working hours or night hours),
(iii) the number of received votes [20] (7 f.: histogram with 7 intervals), and
(iv) the usage of emoticons [11] (150 f.: custom list of 150 emoticons).

Our Additions. Furthermore, we calculate additional idiosyncratic features (8 f.: frequency of grammar mistakes and uppercase word usage per sentence and comment) to also cover more user-specific writing style habits. We additionally measure the amount of whitespace (2 f.: per sentence and message) and the comment length (2 f.) as well as the use of 100 popular hashtags (100 f.), reply functionality (1 f.), quotations (3 f.: usage and placement), and 50 popular link shorteners (50 f.) to cover aspects specific to news comments.

In total, our stylometric feature set consists of 121378 features. While most of these features can be applied for every news site, some specific features, e.g., quotations from previous comments, can only be used if a site implements the functionality. Recall that Doppelgänger Finder applies a PCA for dimension reduction to the feature set, making the effectively used features dependent on the training data. The analysis of feature relevance is beyond the scope of our feasibility study, which is why we leave it to future work.

All these features operate solely on public information, i.e., no private side channel information, such as IP addresses or email addresses, is used. The rational here is that modifications to the comment with the goal to avoid detection of manipulation are unlikely as any changes to the wording can have an impact on the success of the attempted opinion manipulation.

3.2 Applying Doppelgänger Finder to News Comments

To apply Doppelgänger Finder (cf. Section 2.2) to news comments, we first need a set of user accounts (pseudonyms) for our analysis. Then, we need to extract the writing style features (using our stylometric feature set) from a user's comments (with a certain minimum text length). To extract our features, we make use of several comments written under the same account, following the stylometric evaluations of Twitter [8]. We then input these stylometric profiles into Doppelgänger Finder to identify authors with similar writing styles (potential doppelgängers). More specifically, Doppelgänger Finder computes a similarity score for each pair of authors.

To categorize an author pair as either doppelgänger or non-doppelgänger pair, we have to define a threshold to provide a semantic

to our evaluation results, i.e., signal that the respective two author profiles are likely to belong to a single user. Doppelgänger Finder [3] does not provide a structured approach to obtain a threshold. Instead, the authors recommend a manual analysis to determine the threshold, which is tedious and potentially error-prone in scenarios with ambiguous writing styles. In this work, we present a metric to *automatically* derive such a threshold.

3.3 Automated Threshold Metric

As manual analysis to derive the best threshold is very time consuming and prone to errors, we propose an automated metric. The current set of parameters, such as the number of user accounts, the considered text length per account, and the dataset, defines our *target* scenario. Given that the threshold depends on these parameters, we *simulate* scenarios which are similar to our current *target* scenario to obtain a threshold which is generally applicable to scenarios with identical parameters. Here, the overall concept is to rely on artificially splitting user accounts, i.e., constructing known pairs of doppelgänger, to create a new controlled scenario with the same parameters. When splitting a user account (or pseudonym) P, we divide all comments of P randomly into two sets (P_1 and P_2), while making sure that all comments on an article belong to the same set. Instead of evaluating a user account P, we analyze comments written by its respective partitioned pseudonyms P_1 and P_2.

With this approach, we establish ground truth on our *simulated* scenarios (as we know that P_1 and P_2 are doppelgängers), and thus can group all pairs into two classes (doppelgänger and non-doppelgänger pairs). By applying a threshold, we can compare the classifier's output (i.e., similarity scores) to the known grouping of all pairs to assess the classification accuracy. Hence, we can choose the threshold such that we achieve the optimum in terms of precision and recall. Repeating these *simulated* scenarios multiple times (for our *target* scenario) allows us to automatically determine a generally applicable threshold for these particular parameters.

Finally, we can apply this threshold to our *target* scenario (without ground truth). Author pairs with a similarity score (as computed by Doppelgänger Finder) exceeding the threshold are flagged as doppelgängers, while the remaining pairs are considered to be non-doppelgängers. As this approach only provides a *heuristic* for doppelgängers, we recommend to verify this flagging through other means (either automated monitoring or manual analysis).

By crafting a stylometric feature set for news comments and extending Doppelgänger Finder with a metric for automated threshold selection, we provide an approach that allows to check arbitrary sets of author profiles on news sites for doppelgängers, i.e., single individuals publishing comments under different accounts, solely based on information directly derived from the published comments.

4 DATA SET

We evaluate our approach by applying it to detect doppelgängers in user comments on online news sites. Since no public data set of texts published in comment sections exists, we obtain 5 million comments posted in response to 300 k articles in 3 major news papers by crawling. We next describe our crawling approach and the resulting data set.

4.1 Selection Criteria

For our evaluation, we want to target popular news sites that can be of high relevance to opinion manipulation by the ability to reach a large target audience. Hence, we focus on popular news sites (according to their Alexa rank) with an ability for users to post comments. For crawling, we require sites to display comments on their own site, i.e., without embedding third-party components such as *Facebook* or *Disqus*. To study doppelgänger accounts, we require sites to offer a proper authentication mechanism to ensure that all comments are uniquely assigned to a user profile or account number in the data set. Otherwise, displayed author names do not necessarily refer to the same individual and thus might already represent a mixture of writing styles from multiple authors, even without manipulative doppelgängers. Furthermore, dedicated profiles allow us to list all comments published by a single user, easing the collection of a large number of comments for each profile. In our opinion, this constraint does not limit the applicability of our approach because active, connected, and long lasting user accounts are more likely to convince users when spreading misinformation.

Our final selection consists of *the Guardian* (2nd most popular news website in UK), *SPIEGEL ONLINE* (most popular news website in DE), and *ZEIT ONLINE* (5th most popular news website in DE). While selecting pages with different languages helps us to check the language independence of our approach, choosing two pages in the same language allows us to investigate differences between news platforms in the same country.

Only relying on public comment data that is available through crawling forms a minimal set of information that exacerbates the detection problem. That is, platform providers can still enrich our approach by using further side channel information or non-public user account identifiers to map multiple comments to a single pseudonym. Such a proprietary data set is, however, only available to the platform providers themselves, while our approach can be directly applied by the general public.

4.2 Crawling Period

We periodically crawled the selected news sites over a period of two month (July 1, 2017 to August 31, 2017) to collect large comments data sets. This period covers two rounds of United Kingdom's European Union exiting negotiations, which might be a potential target for opinion manipulation and spreading of misinformation.

During our crawling period, we scanned the main page of each news site every 30 s for new articles. For each article, we retrieved all comments at different intervals (after 1 h, 4 h, 12 h, 24 h, 48 h, and 7 d). The motivation here is that our data set will also contain altered comments (moderated, edited, or deleted). Besides, our assumption is that comment sections are most active after an article has initially been published and consequentially, early published comments might cause the most severe damage wrt. opinion manipulation. Finally, after the two month period, we collected the 100 most recent comments per user account (for *the Guardian*, we only processed 62 % of the users as they asked us to cease our crawling activity).

4.3 Data Set Statistics

We retrieved ≈ 5 million authentic comments of nearly 60000 user accounts on over 300000 news articles as summarized in Table 1.

Table 1: The posting behavior of user comments on news depends on the crawled website. As expected (cf. linguostatistics), German language sites on average comprise of longer comments than comments on the English language site *the Guardian*.

Dataset	# Articles	# Users	# Comments	Comments / Article* [⌀]	Users / Article* [⌀]	Comments / User [⌀]	Length / Comment [⌀]
The Guardian	210819	48033	3023674	385.1	85.9	62.9	262.7
SPIEGEL ONLINE	89480	11460	808245	62.2	35.1	70.5	505.8
ZEIT ONLINE	29784	10265	1057175	171.1	67.6	103.0	434.6

* If the article was published in July or August 2017 and the comment section of the article was enabled.

Our data sets also record website-specific features, such as the use of quotations (*the Guardian* and *SPIEGEL ONLINE*), the number of votes (*the Guardian* and *ZEIT ONLINE*), or comment headlines (*SPIEGEL ONLINE*). Besides, we also observe that moderation takes place on the crawled sites (e.g., user accounts being closed or comments being modified/deleted). Although we refrained from assessing the reasons for these actions, misconducts, such as defamation, harassment, spam, or opinion manipulation, are likely reasons.

5 EVALUATION

We evaluate our approach by artificially splitting user accounts (cf. Section 3.3) and thereby *simulating* the presence of doppelgängers in real-world texts. This method is necessary since we have no ground truth data set with labeled doppelgängers. By artificially splitting accounts, we assess the ability of our approach to detect the splitted accounts by the same author, and thereby detecting doppelgängers.

5.1 Simulation of Doppelgängers Approach

Our evaluation approach follows established prior work (e.g., Abbasi et al. [1] or Johansson et al. [16]) to split user accounts into pseudonyms. This way, we introduce *known* doppelgängers into real-world texts without requiring labeled data of doppelgängers. We deliberately decided to experiment on real-world texts instead artificially created texts (e.g., written by paid *Amazon Mechanical Turk* workers [13]) to test our approach on realistic input data.

In our evaluation, we conduct a 3-fold cross-validation for each experiment. First, we randomly select user accounts from our data sets and artificially split them as described previously (acting as pseudonyms). Then, we apply our approach (cf. Section 3) to each fold. Finally, we average the results over all three folds. Note that our results correspond to a lower bound because the random set of user accounts could already contain doppelgänger pairs which we are unaware of due to the missing ground truth.

5.2 Applied Statistical Measures

The traditional definition of precision and recall is not applicable to all evaluation scenarios because the number of included artificial doppelgänger pairs can also be zero (varying number of doppelgängers). Hence, given that we have no true positives, we would always end with a precision and recall of 0 %. An alternative definition of counting the correctly predicted pairs is unsuitable as well because in larger scenarios the non-doppelgänger pairs (increasing quadratically) dominate the result (the number of evaluated accounts only scales linearly), i.e., reducing the use of precision and recall to absurdity. Consequentially, we need a custom definition that is unaffected by the number of included doppelgängers while not being dominated by the majority of non-doppelgänger pairs.

Our following custom definition of the entries in a confusion matrix addresses these issues. Hence, we can compare the results of different evaluations regardless of the adjusted parameter. However, a drawback of this adjustment is that in scenarios with only a few doppelgänger pairs, missed pairs do not result in a low accuracy. In our opinion, this limitation is not an issue because we want to focus on reliably detecting these doppelgänger pairs without introducing incorrectly flagged pairs. Nevertheless, as part of our evaluation, we conduct experiments with varying numbers of doppelgänger pairs to cover different scenarios.

For each pseudonym, we refer to the pairs that the classifier correctly predicts to be a doppelgänger as $\mathrm{cDG_p}$ (correct Doppelgänger pairs). We count the pairs that the classifier incorrectly predicts to be a doppelgänger as $\mathrm{FP_p}$ (False Positive pairs) and the pairs that the classifier fails to predict to be a doppelgänger as $\mathrm{FN_p}$ (False Negative pairs). The lower $_\mathrm{p}$ denotes that we count the number of pairs (each pseudonym is compared to all others). Next, we introduce our mapping from these pair counters, i.e., $\mathrm{cDG_p}$, $\mathrm{FP_p}$, and $\mathrm{FN_p}$, to entries in a confusion matrix for each pseudonym.

$$\mathrm{TP} = \begin{cases} \dfrac{\mathrm{cDG_p}}{\mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p}} & , \text{ if } \mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p} \neq 0, \\ 1 & , \text{ otherwise.} \end{cases}$$

$$\mathrm{FP/FN} = \begin{cases} \dfrac{\mathrm{FP_p/FN_p}}{\mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p}} & , \text{ if } \mathrm{cDG_p} + \mathrm{FP_p} + \mathrm{FN_p} \neq 0, \\ 0 & , \text{ otherwise.} \end{cases}$$

We apply the commonly known definitions of precision and recall to these mappings and present the results as weighted F-score ($F_{0.5}$) that favors precision over recall [27] as we are more interested in correctly detecting pairs of doppelgängers than detecting all doppelgängers (while accepting incorrectly flagged pairs).

5.3 Known Number of Doppelgängers

We begin with evaluating a *baseline scenario* in which we use a-priori knowledge on the number of doppelgängers to correctly set the threshold. We omit this baseline setting in the next section where we do not use a-priori knowledge for a more realistic setting.

Approach. We base our decisions and initial experiments on our *the Guardian* dataset. Our approach requires three different parameters to be set in advance:

(i) the number of accounts that we want to evaluate,
(ii) the number of instances that represent the writing style, and
(iii) similarly, the text length that a single instance consists of.

Overall, we select realistic values that conform to our *the Guardian* dataset. For the number of accounts, we choose a total of 100 pseudonyms since this number exceeds the average number of 85 user accounts commenting on an article (cf. Table 1). Second, this setting still covers 77.67 % articles in our *the Guardian* dataset.

Given the shorter length of news comments compared to forum postings originally used to develop doppelgängers finder [3], we double their choice and select 20 instances per pseudonym. This number of writing style profiles, which we generate per pseudonym, correlates with information that we have available for each pseudonym. The text length per instance determines the quality of a single writing style profile.

Since comments are rather short, we append multiple comments to a longer comment [8] and evaluate two different strategies to identify the optimal decision; we can either take a fixed number of comments, or we can require a minimum number of characters.

Results. We show the doppelgängers detection accuracy as $F_{0.5}$ in Table 2 (row: Complete Scenario) for randomly selected comments with a minimum appended length of 250 to 1000 characters and 2 to 6 randomly selected (and appended) comments. We selected these increments because the average length of a comment on *the Guardian* is 262.70 characters. We refrained from evaluating a single comment because this randomly chosen comment can be significantly shorter than the average. The results highlight a high detection accuracy in every scenario. Further, more textual information improves the classification accuracy.

To increase the applicability (trade-off between accuracy and required comments per pseudonym), we fix the text length per instance at >750 characters because the improvements saturate, i.e., our approach correctly separates doppelgängers and non-doppelgängers for most of evaluated pseudonyms. With these parameter decisions, we require approximately 60 comments for every pseudonym that is part of our evaluation. In our *the Guardian* dataset, this corresponds to more than 99.48 % of all user accounts. Furthermore, we believe that pseudonyms with more published comments have more influence (wrt. opinion manipulation) because they appear more trustworthy than fresh accounts. Hence, we expect that we are able to collect a sufficient number of comments for every pseudonym that is relevant for doppelgängers checks.

When comparing this decision to related work, Afroz et al. [3] train their classifier on at least 4500 words per author. When applying an average word length, such as 5.10 characters [9], we end up with approximately 22950 characters per author. With our setting of 750 characters per instance and 20 instances per author, we end up with a lower bound of only 15000 characters. Consequentially, our *baseline scenario* maps to our setting of shorter targeted texts and provides initial reasonable results to serve as a candidate selection for doppelgänger detection.

Comment Variability. The previous evaluations relied on selecting a fixed, randomly selected set of comments per author and thus did not enable us to assess the influence of variability between different comments on the detection accuracy. To assess the influence of this variability, we now repeat the previous evaluations but multiple times (over 5 runs). In each run, we select different comments from each author. Overall, the variance in the results following randomly chosen comments is marginal and ranges between 4 to 8 incorrectly classified pseudonyms out of 300 (3 folds with 100 pseudonyms each). We thus conclude that the influence of comment variability is insignificant.

Table 2: *Simulated Doppelgängers:* Results for *the Guardian* dataset scenarios with 20 instances per pseudonym and a total of 100 pseudonyms depending on the composition of an instance. We split 50 user accounts into two pseudonyms each and list the $F_{0.5}$-scores measured in percent [%].

Scenario	Length [char.]				Comments [#]		
	250	500	750	1000	2	4	6
●*Complete*	95.11	98.74	99.32	99.73	94.53	98.61	97.13
◐*Partial*	95.92	97.40	99.46	99.57	93.22	97.57	95.30

Restricted News Resorts. Next, we check whether the content of our evaluated comments has an impact on the classification accuracy. We therefore select comments from articles in critical resorts (politics, news, education, and environment) on *the Guardian* that might be more prone to the spreading of misinformation. We list the respective results in Table 2 (row: Partial Scenario). We are unable to observe a definite trend in either direction and therefore, we assume that our approach is feasible for the entire platform of *the Guardian*, irrespective of the comments' content. Consequentially, we conclude that the content of comments has a negligible influence on our proposed approach.

Having analyzed the influence of the text length per instance, we next focus on the remaining two parameters: number of instances and number of pseudonyms. We stick to our *baseline scenario* and only adjust a single parameter at once.

Number of Instances. An increasing number of instances increases the classification accuracy ($F_{0.5}$-score of 99.59 % for 25 instances and 100.00 % for 30 instances). Contrary, too few instances per pseudonym significantly impair the results (e.g., $F_{0.5}$-score of 90.98 % for 10 instances and only 77.43 % for 5 instances). This decision must be made in light of the trade-off between accuracy and required comments per pseudonym. If we analyze more comments per user account, we can create additional stylometric profiles, i.e., differentiate them better from other unrelated profiles. However, we might not have a sufficient number of comments for every account.

Number of Pseudonyms. Adjusting the number of pseudonyms per account provides similar results in scenarios with 150 and 200 pseudonyms ($F_{0.5}$-score of 98.81 % and 98.25 %, respectively). Increasing the parameter even further to 250 and 300 pseudonyms, results in a decline to 97.03 % and 96.76 %. This observation is reasonable because the number of pairs that the doppelgänger detector considers increases quadratically (as the similarity score is calculated for each pair). With an increasing number of considered writing styles, differentiating and matching the writing styles is more challenging. Nevertheless, for reasonable numbers of pseudonyms, our approach identifies most doppelgänger pairs without labeling non-doppelgängers as doppelgängers. Based on our dataset knowledge (an average of 85 user accounts commenting on an article), we can conclude that our proposed approach fits the targeted scenario.

Findings. *The baseline scenario with a-priori knowledge on the number of doppelgängers provides high detection results for realistic comment lengths, the number of comments, and pseudonyms per author.*

Table 3: Comparison of scenarios depending on the number of simulated doppelgänger across all datasets. We list the $F_{0.5}$-scores measured in percent [%].

Dataset	Doppelgänger	None	Single	Random
The Guardian	*matching*	n/a	99.46	98.91
	oblivious	99.46	99.46	99.31
SPIEGEL ONLINE	*matching*	n/a	99.31	98.20
	oblivious	98.38	97.42	98.04
ZEIT ONLINE	*matching*	n/a	99.33	96.71
	oblivious	97.85	98.79	96.91

5.4 Unknown Number of Doppelgängers

We further evaluate a more realistic scenario in which the number of doppelgängers is *unknown* and thus the threshold cannot be set based on the number of doppelgängers. In the following, we show that we achieve comparable results regardless of the number of doppelgängers. We label these experiments as *oblivious* because we do not benefit from knowledge on the expected number of doppelgängers in our simulated scenarios. This adjustment is essential as we are unaware of the number of doppelgängers in a real-world evaluation. We define three new settings:

(i) *None*: our evaluation contains no simulated doppelgänger,
(ii) *Single*: our evaluation contains only a single doppelgänger pair, and
(iii) *Random*: we have between 25 % and 75 % doppelgängers.

For comparison, we label our baseline experiment as *matching* since here the threshold is derived in simulated scenarios where the number of doppelgängers matches the experiment (known number of doppelgängers).

For our *matching* experiments, we determine the thresholds on related scenarios (as conducted before when splitting each user account into two pseudonyms). Consequentially, for the *None* setting, we are unable to determine a threshold without any simulated doppelgänger. A threshold of 1 would always yield perfect results. In *oblivious* experiments, we determine the threshold on a large number of scenarios that equally consist of all our defined settings. Hence, we do not favor a specific quota of doppelgängers when determining the threshold. As stated before, additional scenarios improve the accuracy of the threshold as it depends on the dataset.

In Table 3, we list the respective evaluation results. We again observe a high detection accuracy for all scenarios even if the number of doppelgängers is unknown to the algorithm (oblivious). Compared to the baseline setting in which the number of doppelgängers is known (matching), the detection accuracy is only slightly lower. The slight variation in the *Random* setting results from a varying number of simulated doppelgängers across multiple runs.
Findings. *Our approach is suitable for real-world settings in which the number of doppelgängers is unknown as it still provides a high detection accuracy.*

5.5 Language Dependence

To study the influence of language on the detection accuracy, we now apply our approach to the German data sets (i.e., *SPIEGEL ONLINE* and *ZEIT ONLINE*) by repeating the experiments we presented for our *the Guardian* dataset in Section 5.4.

As listed in Table 3, the results on comments written in German are slightly worse when compared to our *the Guardian* evaluation (regardless of *matching* or *oblivious* experiments). However, the general performance is similar to *the Guardian*. We expect that this slight decrease in accuracy between English and German texts probably results from the longer average text length in German (due to longer words). Repeating the steps that we conducted for *the Guardian* in Section 5.3 to determine the "text length per instance" parameter should improve the presented results marginally. Nevertheless, the results show that our general approach is applicable to German comments without any adjustments.

When comparing the German data sets to each other, we notice minimally worse results on *ZEIT ONLINE*. We believe that this trend results from two aspects. First, the conversations in *ZEIT ONLINE*'s comment sections are more colloquial when compared to *SPIEGEL ONLINE*. Second, as listed in Table 1, the average comment length on *ZEIT ONLINE* is shorter. Hence, the extracted writing style profiles are less accurate. In this case, adjusting the "text length per instance" parameter might improve the achieved results as well.

Adding further languages requires tool support (part-of-speech tagger) and structural information (e.g., word separators) on the target language to properly implement the features.
Findings. *Our approach is able to reliably detect doppelgängers in both languages (English vs. German).*

6 REAL-WORLD IMPRESSIONS

We now apply our approach in a real-world setting (without ground truth). More specifically, we conduct tests similar to what could actually be performed by news sites to detect online manipulation; we manually inspect the doppelgängers detected by applying our approach. Even in the absence of ground truth, the motivation of our anecdotal analysis is to get a first intuition on whether the detected pairs of doppelgängers are realistic.

To limit the manual effort required for checking detected doppelgängers, we restrict our analysis to 10 comment sections per website and only compare user accounts that comment on the same article. For each article, we compare a random subset of 100 user accounts to each other. We use the same parameters as in our baseline scenario (cf. Section 5.3): 20 instances per user and at least 750 characters per comment. For each data set, we use the respective threshold (t) as derived in Section 5.4.
The Guardian. In the comment section of an article demanding politicians to investigate the funding of (Islamic) extremists in Great Britain, our approach detects two accounts with a similarity score of 0.0218 (t = 0.0140). The two accounts were registered within half a year (2012/2013) and published more than 3200 comments. Manual inspection indicates that writing styles and behaviors, e.g., preferred articles, overlap. Furthermore, both accounts oppose Brexit.

Commenting users of an article where a man is reportedly urged to take a citizenship test despite born and living in Great Britain for

his whole life lead to a pair of potential doppelgängers with a similarity score of 0.0161 (t = 0.0140). Both accounts have about 210 comments but were registered in different years (2010 vs. 2016). Manual investigation reveals that both authors write long comments. However, while one frequently inserts line breaks, the other account publishes long paragraphs. Hence, without additional information, assessing whether these two accounts are indeed doppelgängers is difficult.

SPIEGEL ONLINE. Analyzing user accounts commenting on an article about an anti-terror demonstration in Cologne, Germany, our approach identifies a doppelgänger pair with a similarity score of 0.0220 (t = 0.0160). While one of the accounts has more than 1700 comments and was registered in 2008, the other one is only rarely used for commenting (around 80 comments) since its registration in 2015. Both accounts mainly comment on articles about politics; the account with more comments is also active in resorts that are less frequently accessed, e.g., culture, panorama, and health. An additional manual analysis confirms that the two writing styles are very similar. Furthermore, both accounts frequently use comment titles, an optional feature only offered by *SPIEGEL ONLINE*.

ZEIT ONLINE. In the comments to an article on the European Union's Brexit demands, our approach uncovers a (potential) doppelgänger pair with a similarity score of 0.4445 (t = 0.0120). Manual analysis shows that both accounts publish English-language comments on a German-language news site. As English writing styles strongly differ from German writing styles, such a result is not surprising. Noticing clear differences in the command of the English language, we strongly believe that these authors are not real doppelgängers.

The same article reveals an interesting situation with two additional, overlapping pairs of (potential) doppelgängers. Authors A and B have a similarity score of 0.0203 (t = 0.0120), while A and C have a similarity score of 0.0177. The similarity score of the third pair, i.e., B and C, is slightly below the threshold with 0.0111. To further investigate these potential doppelgängers, we cannot access the registration dates (not publicly provided by *ZEIT ONLINE*), and hence have to find other indicators to infer their actual relationship. Considering their activity periods, we find that A is active throughout the day, including night hours. A similar observation holds for B, while C is only active during typical office hours. Likewise, we observe that the writing styles of A and B are more alike than their writing styles compared to C. As supported by the similarity scores, we thus believe that A and B indeed might be doppelgängers, whereas A and C as well as B and C are likely not doppelgängers.

Overall, our anecdotal analysis detected some interesting pairs of *potential* doppelgängers. Our manual investigation is, however, unable to assess the pairs with certainty given the lack of ground truth. We contacted the respective German news sites to inquire about these potential doppelgängers, but did not receive replies. Interestingly, all but one of the detected user accounts published comments which are positioned at a prominent location directly below the text of corresponding news articles. These prominent comments are additionally indexed by the Google search engine, further increasing their potential impact on public opinion. A relevant direction for future work is therefore to apply our approach to a dataset with known doppelgängers to evaluate its performance against ground truth data. Since such a data set is unavailable to us, we leave this analysis for future work.

7 CONCLUSION

In this paper, we presented an automated approach to detect doppelgängers in comment sections of news websites based on stylometric information. In contrast to approaches that rely on technical metadata (such as IP addresses or browser identifiers)—which can be easily spoofed—our doppelgänger detection makes use of comparison of authors' "fingerprints" derived from their writing style. By using artificially splitted accounts from real-world data sets, we showed that our method is efficient in linking them together for different practical application scenarios. Our preliminary results of an application in the wild showed indications for doppelgängers in user comments of popular online news websites. These findings need to be further investigated and manually inspected to draw final conclusions about the limits of a practical applicability of our method. With our work we want to encourage the community of researchers to further investigate this topic as public opinion manipulation is a serous threat to our society with a potential to cause a severe negative influence on different spheres of our life, even in established democracies.

ACKNOWLEDGMENT

The authors would like to thank Afroz et al. [3] for making their *Doppelgänger Finder* implementation publicly available. This work has been funded by the Excellence Initiative of the German federal and state governments.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM TOIS* 26, 2 (2008).
[2] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *IEEE S&P*.
[3] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger Finder: Taking Stylometry to the Underground. In *IEEE S&P*.
[4] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236.
[5] Mishari Almishari, Dali Kaafar, Gene Tsudik, and Ekin Oguz. 2014. Stylometric Linkability of Tweets. In *WPES*.
[6] Mishari Almishari and Gene Tsudik. 2012. Exploring Linkability of User Reviews. In *ESORICS*.
[7] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).
[8] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. 2013. Stylometric Analysis for Authorship Attribution on Twitter. In *BDA*.
[9] Patrick Blogamundo: Hall. 2007 (accessed January 8, 2018). *Languages by Average word length.* https://web.archive.org/web/20090823115022/http://blogamundo.net:80/lab/wordlengths (original website inactive).
[10] Pew Research Center. 2018 (accessed January 4, 2019). *Digital News Fact Sheet.* http://www.journalism.org/fact-sheet/digital-news/.
[11] Marco Cristani, Giorgio Roffo, Cristina Segalin, Loris Bazzani, Alessandro Vinciarelli, and Vittorio Murino. 2012. Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging. In *ACM MM*.
[12] Chris Elliott. 2014 (accessed December 14, 2018). *The readers' editor on… pro-Russia trolling below the line on Ukraine stories.* https://gu.com/p/3pvee.
[13] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *ACL*.
[14] Andrew J. Flanagin and Miriam J. Metzger. 2000. Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000).
[15] Elle Hunt. 2017 (accessed December 14, 2018). Facebook purges tens of thousands of fake accounts to combat spam ring. (2017 (accessed December 14, 2018)). https://gu.com/p/6afbt.

[16] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. 2013. Detecting Multiple Aliases in Social Media. In *IEEE/ACM ASONAM*.

[17] Fredrik Johansson, Lisa Kaati, and Amendra Shrestha. 2015. Timeprints for Identifying Social Media Users with Multiple Aliases. *Security Informatics* 4, 1 (2015).

[18] Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *WWW*.

[19] Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu, and Jidong Shao. 2015. Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns. In *AAAI ICWSM*.

[20] Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015. Exposing Paid Opinion Manipulation Trolls. In *ACL RANLP*.

[21] Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi D. Georgiev, and Ivan Kolev Koychev. 2018. The dark side of news community forums: opinion manipulation trolls. *Internet Research* 28, 5 (2018).

[22] Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. In *ACL*.

[23] Eni Mustafaraj and Panagiotis Takis Metaxas. 2017. The Fake News Spreading Plague: Was it Preventable?. In *ACM WebSci*.

[24] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the Feasibility of Internet-Scale Author Identification. In *IEEE S&P*.

[25] Sergey Sanovich. 2018. *Russia: The Origins of Digital Misinformation*.

[26] Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship Attribution of Micro-Messages. In *ACL EMNLP*.

[27] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI*.

[28] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2013. A Case Study of Sockpuppet Detection in Wikipedia. In *ACS LASM*.

[29] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60, 3 (2009).

[30] Jon Swaine. 2018 (accessed December 14, 2018). Twitter admits far more Russian bots posted on election than it had disclosed. (2018 (accessed December 14, 2018)). https://gu.com/p/8vp2h.

[31] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018).

[32] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2019. Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter. (2019). arXiv:1901.05997

[33] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM IMC*.

[34] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *WWW Companion*.

[35] Savvas Zannettou, Tristan Caulfield, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2018. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. (2018). arXiv:1811.03130

[36] Xueling Zheng, Yiu Ming Lai, Ka Po Chow, Lucas C.K̃. Hui, and S.M̃. Yiu. 2011. Sockpuppet Detection in Online Discussion Forums. In *IIH-MSP*.

Characterizing Activity on the Deep and Dark Web

Nazgol Tavabi[1],Nathan Bartley[1],Andrés Abeliuk[1],Sandeep Soni[2],Emilio Ferrara[1],Kristina Lerman[1]
1. University of Southern California, Information Sciences Institute
2. Georgia Institute of Technology

ABSTRACT

The deep and darkweb (d2web) refers to limited access web sites that require registration, authentication, or more complex encryption protocols to access them. These web sites serve as hubs for a variety of illicit activities: to trade drugs, stolen user credentials, hacking tools, and to coordinate attacks and manipulation campaigns. Despite its importance to cyber crime, the d2web has not been systematically investigated. In this paper, we study a large corpus of messages posted to 80 d2web forums over a period of more than a year. We identify topics of discussion using LDA and use a non-parametric HMM to model the evolution of topics across forums. Then, we examine the dynamic patterns of discussion and identify forums with similar patterns. We show that our approach surfaces hidden similarities across different forums and can help identify anomalous events in this rich, heterogeneous data.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; **Exploratory data analysis**; *Bayesian nonparametric models*; *Cluster analysis*; • **Information systems** → **Deep web**; *Clustering*; *Web crawling*; • **Security and privacy** → *Social aspects of security and privacy*.

KEYWORDS

Darkweb; Deepweb; D2web; Cyber Crime; Cyber Security; LDA; Non-Parametric HMM; Beta Process; multivariate Time Series; Cluster Time Series

ACM Reference Format:
Nazgol Tavabi[1],Nathan Bartley[1],Andrés Abeliuk[1],Sandeep Soni[2],Emilio Ferrara[1],Kristina Lerman[1]. 2019. Characterizing Activity on the Deep and Dark Web. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316502

1 INTRODUCTION

The web that most people are familiar with—the open and searchable internet of social media sites, online merchants, newspapers and the like—represents just a tiny fraction of the internet. Much larger portions of internet data remain buried within the "deepweb" [18], a term that refers to private corporate intranets and databases, dynamically-generated web pages, and limited access content, such as online academic journals. While only some of the content is hidden behind encrypted protocols in .onion domains, access to the remaining "open" content is generally restricted, requiring registration and authentication [26]. Some portion of the deepweb, also known as the "darkweb," serves as a hub for all kinds of illicit activities. Malicious actors congregate virtually on dark web forums and marketplaces to trade illicit information, goods and services, including ransomware, exploits, hacking tools, stolen media, user credentials, fake ids, prescription medicines and illegal drugs. In addition to serving as a marketplace for these goods, the deep and dark web provides a venue for malicious actors to coordinate cyber attacks [25] and terrorist activity [7]. The growing popularity of the marketplaces within the deepweb can be attributed to the elimination of the risk of violence since there is limited, if any, physical interaction between the buyers and the sellers. Another reason is the use of encrypted protocols to preserve anonymity, encouraging people to express themselves without the risk of getting caught by law enforcement nor being censored by the moderators of a web site [27].

Given the threat posed by these malicious actors, observing their activities on the deep and dark web (d2web) may provide valuable clues both for anticipating and preventing cyber attacks as well as mitigating the fallout from data breaches. However, picking out useful signals in the vast, dynamic and heterogeneous environment of the d2web can be challenging.

In this paper, we use Latent Dirichlet Allocation (LDA) [5] to analyze a large heterogeneous text corpus from the d2web. To understand the dynamics of discussions, we use a non-parametric hidden Markov model [10]—the *Beta Process HMM*. In this approach every forum is represented as a multivariate time series, where variables are the topics found by LDA, and is fed into a Beta Process HMM (BP-HMM). This BP-HMM then finds the shared states among forums, where each state is a distribution over topics. This helps track discussions on different forums and identify anomalous behavior or important events. This approach can also be used to find forums relevant to a specific subject, i.e., forums or time periods within forums where users discuss specific topics, such as hacking techniques and cyber security related issues. This method can also cluster forums into meaningful groups.

We test this framework on data consisting of posts published on 80 D2web forums from 2016 to mid 2017, on topics such as exploits and hacking techniques, selling prescription and non-prescription drugs, and creating fake ids. Overall this paper makes the following contributions:

- Using LDA, we characterize the content of 80 d2web forums where illicit activities are discussed.
- We describe an application of a non-parametric HMM model to learn the forum's shared topic dynamics based on the results obtained from LDA.
- We use the learned shared behaviors as a compact representation of these time series to cluster forums into groups with

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316502

distinct characteristics and analyze learned latent behavioral structures to gain more insight into this data.
- We rank forums based on how their topics of discussions are likely to change.
- Finally, we present case studies revealing how our approach can be used to to identify anomalous activity in d2web data.

2 RELATED WORK

For an overview of the darkweb and deepweb and the challenges these sites pose for researchers and for law enforcement, please see [11, 18]. Researchers have leveraged d2web content in specific applications, but to the best of our knowledge, few have attempted to systematically characterize the topics and dynamics of d2web discussions. For example, Xu et al. [31] analyzed the topology and structure of darkweb networks. Soska et al. [27] analyzed 16 different marketplaces to extract information on goods being sold and money transactions, then trained classifiers on them. Tavabi et al. [28] and Almukaynizi et al. [2] used features from d2web discussions to predict which new vulnerability will be exploited. Goyal et al. [13] and Deb et al. [8] used the frequency of important cyber security-related keywords and the sentiment of d2web posts respectively to predict cyber attacks. Along similar lines, [6] utilized sentiment to analyze communications of extremists on the darkweb and [1] developed a bilingual (English and Arabic) sentiment analysis lexicon for cyber security and radicalism on darkweb forums.

In this paper we propose a statistical approach to analyze discussions of multiple malicious forums by extracting their similarities and their differences through learning hidden Markov models on topic weights obtained from LDA. Latent Dirichlet Allocation (LDA) proposed by Blei et al. [5] is a powerful ubiquitous tool which allows documents to be explained by latent topics. LDA has shown to be effective in the analysis of the darkweb [19, 24]. However, the dynamics of darkweb topics, which can give deeper insight into this data, has not been analyzed previously. Rios et al. [24] used LDA to detect overlapping communities in the darkweb and L'huillier et al. [19] applied LDA to extract key members in darkweb forums. Extensions to LDA, like Dynamic Topic Models [4], have been designed to model evolution of topics across time. However, such models have strong memory requirements making them impractical for a large heterogeneous corpus like the d2web. An alternative approach is performing LDA on the entire corpus and observing fluctuation of weights across time, similar to approaches set forth by [20, 30] and many others. Variants of LDA which combine Hidden Markov Models have also been proposed. The approach proposed by Griffiths et al. [14] models both semantic and syntactic dependencies of documents by defining one state in HMM as the semantic state modeled by LDA and other states as syntactic components. Gruber et al. [15] modeled the relationship between words in a document with an HMM. Their model assumes words in the same sentence have the same topic, and successive sentences are more likely to have the same topics, where in the original LDA paper words in a document are assumed to be independent and documents are modeled as bag of words. These variations were proposed to better capture LDA topics, although Hidden Markov models can also be used to identify shifts and variations in the topics discussed, which is the approach of this work. HMM states can help us recognize important events and anomalies.

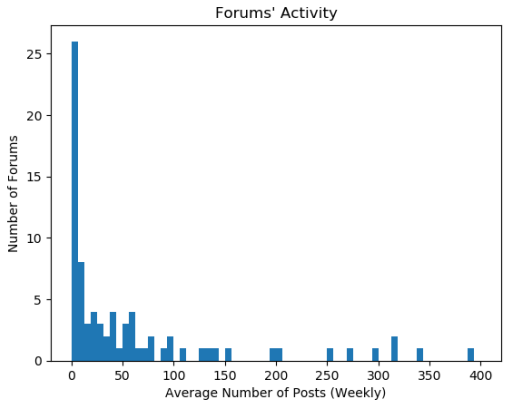


Figure 1: Histogram of activity level in different forums.

3 METHODS

3.1 D2Web Data Collection

The d2web data we use in the study was collected using the crawling infrastructure described in [22, 25]. This crawling infrastructure uses anonymization protocols, such as Tor and I2P, to access darkweb sites, and handles authentication to access non-indexed deepweb sites on the Internet. The infrastructure includes lightweight crawlers and parsers that are focused on specific sites related to malicious hacking and/or online financial fraud. These sites represent forums and discussion boards where people mostly discuss cyber crime and fraud, although other illicit activities are discussed as well, such as the sale of drugs and other stolen goods. There are also a handful of forums crawled that are on the clearnet but which are mostly white hat (i.e., involved with ethical hacking and/or professional cybersecurity). We include these forums in our analysis primarily to help identify other forums that might discuss similar topics. In all, crawlers scraped data from over 250 d2web forums. The most common languages in which the posts were written were English (accounting for 37.8% of all posts), Russian (22.4% of all posts), and Chinese (15.4% of all posts). Other languages, such as Spanish, Arabic, and Turkish were less frequent, each accounting for less than 7% of the posts. For this analysis we only focus on English posts, though the same structure could be used for multiple languages. Filtering out the non-English posts brings the number of forums down to 155. We pre-processed the posts using NLTK [3], SpaCy [17], and scikit-learn to remove stopwords, tokenize each post, and filter tokens by post frequency to remove frequent words. This gives us a corpus of 1.33 million posts.

3.2 Modeling Topics of Discussion

We applied a popular statistical technique known as Latent Dirichlet Allocation (LDA) [5] to learn the topics of the English-language posts. LDA is used to decompose documents into latent topics, where each topic is a distribution over words, intended to capture the semantic content of documents. In this model each document

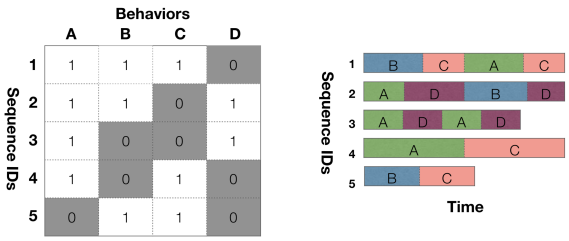
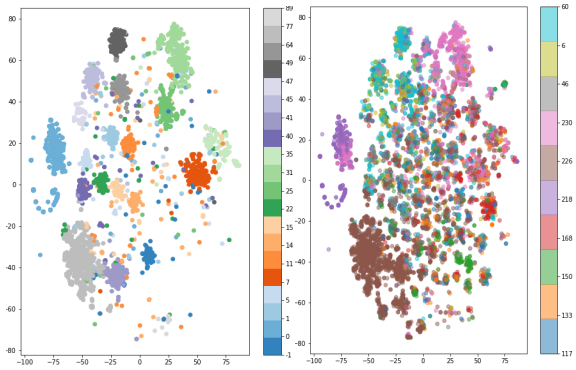


Figure 2: Illustration of the model. Left image is matrix F. Based on this matrix time series 1 exhibits states A,B and C but doesn't exhibit state D. The right image shows segmentation of time series with hidden states



(a) topics (b) forums

Figure 3: Visualization of the d2web discussions in May 2016 using t-SNE. Each dot represents a post, with its color representing (a) the topic or (b) forum to which the post belongs. The top 20 topics learned by the 100-topic LDA are shown, and the post was assigned to the highest probability topic.

is treated as a bag-of-words. Once we learn the model we can represent documents as distributions over a fixed number of topics. Doing so gives us low-dimensional representations of documents.

In this paper we use the Gensim implementation of LDA [23] to learn a model with 100 topics. We train the model on all 1.33 million documents to learn the most informative topics. We tested with 50, 100, and 200 topics respectively, and found that 100 topics results in the most coherent and relevant topics. To examine the dynamics, we focused on time period of 2016 until mid September 2017 which has the best coverage in our data. We only looked at forums with at least one month of activity and more than 100 posts overall, which reduced our data set to 80 forums (and approximately 482 thousand posts). Figure 1 shows the level of activity in these 80 forums. The activity is highly heterogeneous, with some forums seeing hundreds of posts per week, and other forums showing little activity.

3.3 Modeling Dynamics of Activity

Hidden Markov Models (HMMs) have been used extensively in modeling dynamic processes and time series in a variety of applications. These generative models segment time series into a predefined number of latent states and learn transition rates between them.

When modeling multiple dynamic processes with HMMs, it is useful to represent them with global states that are shared between these time series, rather than modeling each dynamic process independently and then learning the mapping between states. Joint modeling facilitates comparison of different processes and learns more generalizable models. It is also convenient to work with a non-parametric model that does not fix the number of states a priori.

In this paper, to model activity within the d2web, we describe each forum as a time series of topic vectors representing discussions. We use a generative model proposed by [9, 10], called *Beta Process HMM* (BP-HMM), to identify latent states shared by different time series. Based on the proposed model, different time series are described by a subset of shared latent states. The states are represented by a binary matrix F, where $F_{ij} = 1$ means time series i is associated with state j. An example of F matrix is shown in the left panel of Figure 2. Given matrix F, each time series is modeled as a separate hidden Markov model with the states it exhibits. Each global state is modeled using a multivariate Gaussian distribution, when a time series is in state x, its data is sampled from a Gaussian distribution with mean vector μ_x and covariance matrix Σ_x.

Since the number of such states in the data is not known a priori, the Beta process is used [16, 29] as a prior on matrix F. A Beta process allows for infinite number of states but encourages sparse representations. Consider, as an example, a model with K states. Each state (column of matrix F) is modeled by a Bernoulli random variable whose parameter is obtained from a Beta distribution (Beta Bernoulli process), i.e.,

$$\begin{aligned} \theta_k &\sim \text{Beta}(\alpha/k, 1), k = 1, \cdots, K \\ F_{nk} &\sim \text{Bernoulli}(\theta_k), n = 1, \cdots, N \end{aligned} \tag{1}$$

The underlying distribution when this process is extended to an infinite number of states, as K tends to infinity, is the Beta process. This process is also known as the *Indian Buffet Process* [12, 29] which can be best understood with the following metaphor involving a sequence of customers (time series) selecting dishes (states) from an infinitely large buffet. The first customer enters the buffet and selects servings from $Poisson(\alpha)$ number of dishes. The n-th customer selects dish k with probability m_k/n, where m_k is the popularity of the dish, yielding the so-called "rich-get-richer" effect, and $Poisson(\alpha/n)$ new dishes.

With this approach, the number of states can grow arbitrarily with the size n of the dataset: in other words, the number of states increases if the data cannot be faithfully represented with the already defined states. However, the probability of adding new states decreases according to a $Poisson(\alpha/n)$. Finally, the distribution generated by the Indian Buffet Process is independent of the order of the customers (time series). For posterior computations based on MCMC algorithms, the original work is referenced [9, 10].

3.4 Clustering

To define a similarity measure between two HMMs, one could measure the probability of their state sequences having been generated by the same process. Since each time series is associated with a

Topic (Manually Labeled)	Top 10 Keywords
Vending	
1. Locations	checked, live, united states, unknown, california, carolina, south, ca, nj, new
2. Money	money, people, make, pay, want, free, buy, just, like, sell
3. Pharmaceuticals	buy, online, prescription, cheap, cod, xanax, delivery, overnight, order, day
4. Banking	card, bank, credit, cards, paypal, account, business, debit, accounts, gift
5. Fake IDs	fake, real, id, original, high, english, license, quality, registered, passports
6. Purchase details	order, vendor, days, sent, orders, ordered, package, received, shipped, just
7. LSD	like, just, quote, lsd, really, good, feel, know, tabs, experience
8. Cryptocurrency	bitcoin, btc, wallet, address, send, coins, bitcoins, transaction, sent, account
9. Marijuana	like, good, got, weed, time, bit, high, great, nice, low
10. Markets	market, vendor, vendors, dream, alphabay, markets, scam, ab, hansa, escrow
11. Narcotics	good, cocaine, quality, best, vendor, product, mdma, coke, free, order
Security	
12. Malware	virus, scan, antivirus, file, malware, clean, av, security, download, detected
13. Botnets	bot, attack, malware, used, domain, ddos, botnet, irc, hosting, attacks
14. Windows	windows, build, microsoft, xp, vista, beta, server, ms, longhorn, version
15. Social hacking	email, send, rat, stealer, message, keylogger, mail, facebook, crypter, download
16. Law enforcement	police, law, drug, drugs, enforcement, according, dark, illegal, said, darknet
17. Hacking tutorial	learn, know, want, good, learning, start, like, knowledge, programming, hacking
18. Carding	transfer, dumps, info, sell, cvv, good, track, balance, bank, uk
19. Web Vulnerabilities	web, sql, php, injection, exploit, code, server, site, script, page
20. OS Code	process, code, dll, memory, address, api, function, module, use, hook
21. Network Hacking	network, connect, wifi, wireless, ip, internet, router, connected, pineapple, fon
22. Security	information, data, security, software, used, access, user, users, network, application
23. Proxy	use, tor, using, vpn, internet, proxy, browser, ip, web, access
24. Mobile phones	phone, android, phones, samsung, pixel, battery, note, camera, better, google
25. Update	install, installed, download, just, update, need, use, installing, using, try
Gaming	
26. Gaming Source Code	end, local, return, function, false, mod, script, item, nil, damage
27. Torrents	torrent, quote, download, upload, forget, left, feedback, like, plz, 720p
28. Gameplay	game, complete, level, win, play, kill, mode, team, player, single
29. Games	game, games, new, play, like, xbox, ps4, sony, playstation, console
30. PlayStation Vita	vita, ps, psp, game, firmware, exploit, games, sony, custom, psn
31. Emulators	game, games, vita, version, plugin, homebrew, psvita, emulator, use, play
32. Hacking Consoles	ps3, games, play, tutorial, game, cfw, console, psn, use, need
Other	
33. Contact	contact, pm, need, icq, want, send, add, rue, interested, na
34. Thanks	thanks, thank, man, lot, sharing, bro, thx, share, mate, nice

Table 1: Example topics in the 100-topic LDA model.Topics are labeled (in the first column) manually for convenience.

distinct generative process, we measure two state sequences' similarity as the likelihood that seq_i was generated by the process that gave rise to seq_j, and the likelihood that seq_j was generated by the process giving rise to seq_i. We average the two likelihoods to symmetrize the similarity measure.

$$\forall_{i,j}\text{Sim}(i,j) = \frac{p(\text{seq}_i|T_j) + p(\text{seq}_j|T_i)}{2} \tag{2}$$

The likelihood $p(\text{seq}_i|T_j)$ is computed using the learned transition matrix T_i and Markov process assumption. In transition matrix T_i, which is a square matrix with dimensions equal to the number of states, entry $T_i(m,n)$ gives the probability that time series i transitions from state m to state n. Matrix T_i is stochastic, with the sum of entries in each row equal to 1. Once the similarity between two HMMs is defined, we can perform a number of operations, including clustering similar time series together. For example, we use hierarchical agglomerative clustering method to automatically group forums (represented by their time series) with similar discussions.

4 RESULTS

4.1 Topic Analysis

In this section we explore topics learned by Latent Dirichlet Allocation (LDA) by looking at their most significant words. There are a wide variety of topics covered in our dataset. Table 1 highlights some of the topics learned by the 100-topic LDA model by showing the most significant words associated with each topic. We visualized the topics using t-distributed Stochastic Neighbor Embedding

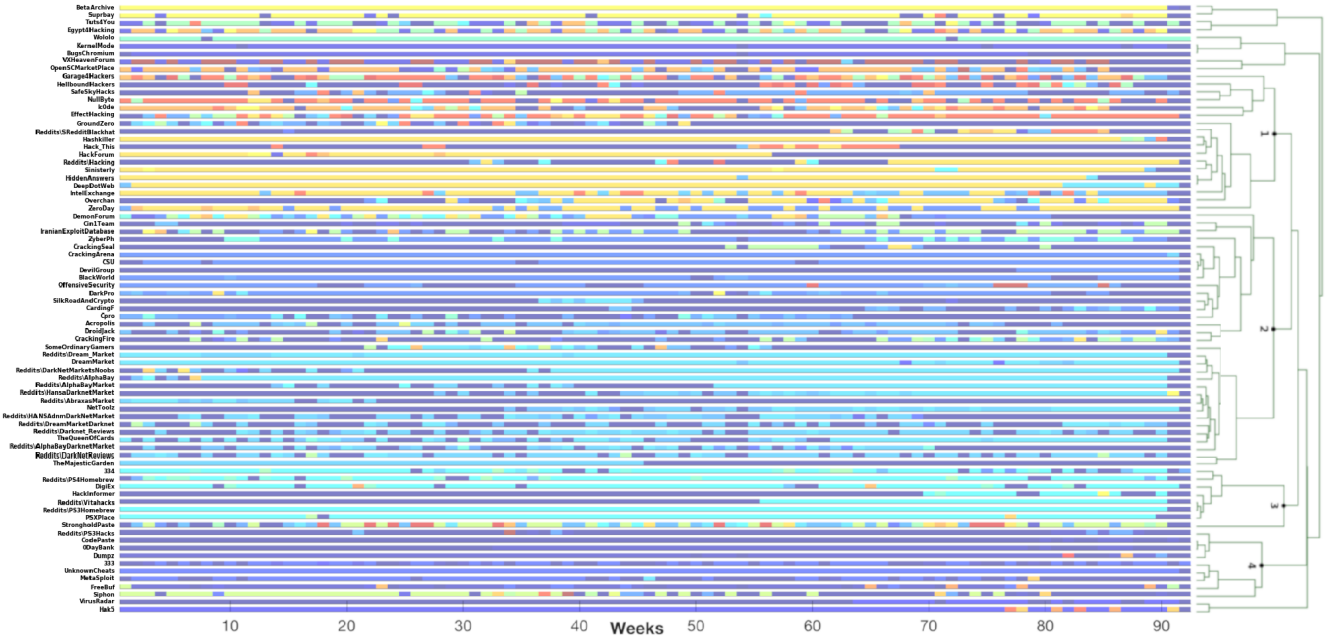


Figure 4: State sequences of forums (each color represents a state) and Dendrogram showing the similarity of forums based on their learned states.

(t-SNE) [21] in Figure 3(a). Each dot on the plot represents a post, colored by its most important topic. Alternatively, we can color the posts in the t-SNE space by the forum on which they were posted (Figure 3(b)). The clustering of posts in the forum space suggests that topics are highly concentrated around specific forums.

4.2 Forum Dynamics

To examine the dynamics of topics on d2web forums, we represent each forum as a time series of topic vectors learned by the 100-topic LDA model. Our unit of time in this analysis is a week; to generate a forum's vector we average the topic vectors of all posts submitted to the forum over the course of a week. The time series of weekly topic vectors were used to both learn HMM states and compute cross-entropy (cross-entropy is discussed further in 4.3).

After training the BP-HMM model on weekly topic distributions of forums the model learned 28 states, i.e., 28 different topic distributions. We clustered the forums according to the similarity of their learned states using the method described in Section 3.4. Figure 4 shows the resulting dendrogram and also the sequences of learned states for each forum. Each line in the figure represents a forum, and different states are represented by different colors. Transitions between states are visible in places the colors alternate.

The clustering results show that the method is able to cluster forums into meaningful groups. Next we examine a few of the main clusters:

Cluster 1 mostly contains forums discussing cyber hacking, including HackForum, GroundZero, ZeroDay, DeepDotWeb, SafeSkyHacks. The two subgroups in this cluster differ mainly in their levels of activity. Forums in the first subgroup are less active (5 posts in a week on average), which is why their average weekly topic vector is more sensitive and their corresponding HMM changes state more frequently. In the active group (more than 50 posts in a week on average) the most common state is the yellow state, which corresponds to activation of the following topics (described in Table 1): 16.Law enforcement, 23.Proxy and 17.Hacking tutorial.

Cluster 2 mostly contains dark web marketplaces such as Abraxas Market, AlphaBay, Dream Market, Hansa Market and BlackWorld, as well as forums dedicated to their reviews. This cluster is also divided into two main subgroups: in the first subgroup the dark blue state dominates representing high activity of discussions regarding topics 1.Locations (which is mostly concerned with the sale of proxy servers), 33.Contact, 34.Thanks and 4.Banking. The second subgroup, depicted with light blue states, corresponds to the topics 6.Purchase details, 10.Markets, 8.Cryptocurrency and 11.Narcotics. Based on the clustering, one can characterize forums in the first subgroup as mostly selling proxy servers and sharing information about other marketplaces, while forums in the second subgroup are more involved in selling drugs. In section 4.4.2 we take a deeper look into forums in the first subgroup.

Cluster 3 is made up of forums related to hacking Playstation video game consoles where the most prominent state in this cluster is the Cyan state. The most active topics in this state are 32.Hacking Consoles and 25.Update.

Cluster 4 (as well as the two forums adjacent) contains forums which are predominantly focused on white hat hacking. Notable forums include Metasploit and Hak5. While 0daybank and FreeBuf are related, they are mostly in Chinese, hence their more active topics have many non-English tokens and are hard to interpret.

With the state sequences obtained from our BP-HMM model, we can track forums' discussions. State transitions indicate a significant change in discussions and could represent an event. However, as shown in Figure 4, some forums change states more frequently, thus their transitions might have less significance. In order to be able to recognize significant transitions, we describe the volatility measure, a forum's likelihood to drastically change its own topic distribution. As each forum is characterized by its learned transition matrix over the global states, we compute a forum's volatility by adding the off diagonal elements of its transition matrix, the probability of changing states. Since we are interested in finding variations in topics discussed rather than the activity of forums, the probabilities of the state corresponding to 0 posts (i.e., no data) were not taken into account. Also to validate the results obtained, a similar volatility measure was computed with cross entropy and is described in section 4.3. Table 2 shows a list of forums with high and low volatility computed via both methods.

Using the described volatility measure with the HMM, we find that the most volatile forums with at least 10 posts per week (on average) are OpenSC Marketplace, Stronghold Paste and Demon Forum and that the least volatile forums are BugsChromium, CSU, and the subreddit PS3Homebrew. An estimate of forum's volatility or lack thereof is also apparent from its state sequence in Figure 4. Stronghold Paste is an onion website similar to Pastebin and covers different topics and hence different states. However, there are two main states it oscillates between: in one of them hacking and cyber security topics (described by topic 19.Web Vulnerabilities) are more prominent, and in the other one, topic 8.Cryptocurrencies has high activation. These results show state transitions in forums like Stronghold Paste have a low probability of being indicative of an event. However transitions in forums like BugsChromium or CSU might be of more interest.

4.3 Topic Dynamics

Since our HMM segments forum dynamics into discrete states, it is less sensitive to noise observed in the data. On the other hand it might not be able to capture small meaningful changes or trends. Hence, we also compute cross entropy of forums as a measure of dispersion over time to validate the results obtained with our non-parametric HMM model. To compute the cross-entropy we use the following formula where Q is the topic distribution for a forum averaged over the entire timespan and P is the topic distribution in a unit of time.

$$H(p, q) = E_P[-\log_2 Q(x)] \tag{3}$$

The majority of the forums have cross-entropy values in the same range. The forum with the lowest average cross-entropy (and therefore lowest volatility) is CSU, a forum primarily concerned with credit card dumps, and as such makes sense that the forum would focus on very few topics of discussion. On the other hand, forums like CodePaste and DroidJack have some of the highest values of cross-entropy, which suggests that the topics of discussion are dispersed and change more often in these forums. Table 2 shows forums with lowest and highest average values of cross-entropy, used as a measure of volatility. This is consistent with the result we got from the HMM model, in the sense that forums with low or high volatility based on the HMM measure also appear in the bottom or top of the ranking based on the cross entropy measure. Results show that large and active forums have wide-ranging discussions on diverse topics however their average topic distribution is usually consistent over time. Forums focused on specific topics like CSU also tend to have low volatility. On the other end of the spectrum there are forums with medium or low activity with discussions spanning a wide range of topics like Stronghold Paste and CodePaste.

HMM-ranked volatility	Cross-entropy-ranked volatility
1. GroundZero	CodePaste
2. OpenSCMarketPlace	DroidJack
3. StrongholdPaste	EffectHacking
4. DemonForum	DemonForum
5. EffectHacking	Overchan
6. CrackingFire	HellboundHackers
7. Overchan	CardingF
8. NullByte	HackForum
9. Siphon	Reddits\ Hacking
...	...
71. UnknownCheats	TheMajesticGarden
72. DevilGroup	Dumpz
73. KernelMode	KernelMode
74. VirusRadar	BlackWorld
75. Reddits\Vitahacks	MetaSploit
76. BetaArchive	0DayBank
77. TheMajesticGarden	Reddits\DarkNetReviews
78. Reddits\PS3Homebrew	Wololo
79. CSU	VirusRadar
80. BugsChromium	CSU

Table 2: Volatility computed via HMM and via Cross-entropy. Ranked highest volatility to lowest.

4.4 Case Studies

4.4.1 Prescription Drugs. In this section we give an example of how this framework could be used to study d2web discussions. In the search for anomalies using the results obtained by our HMM we observed a rare state, exhibited only on 3 first weeks of June 2017 and first week of August 2017 by forum OffensiveSecurity where topic 3.Pharmaceuticals has its highest probability. OffensiveSecurity is considered a low volatility forum based on both measures computed in Table 2 which makes this transition more of interest.

By looking at the posts published on this forum and on the aforementioned dates we retrieved similar posts with variations in the names of the drugs being advertised. An excerpt from one of the posts is as follows: "...buy lynoral cheap buy generic femara buy modafinil online uk can i buy qsymia online buy cytotec online us buy lumigan online canada buy 25 mg lyrica buy vibramycin florida buy diflucan without buy synthroid online next day delivery buy xanax online us where to buy generic qsymia buy adderall no prescription buy cytotec in europe..."

This analysis shows a high and anomalous volume of advertisements for prescription drugs in the specified dates which suggests some precipitating event that merits further investigation.

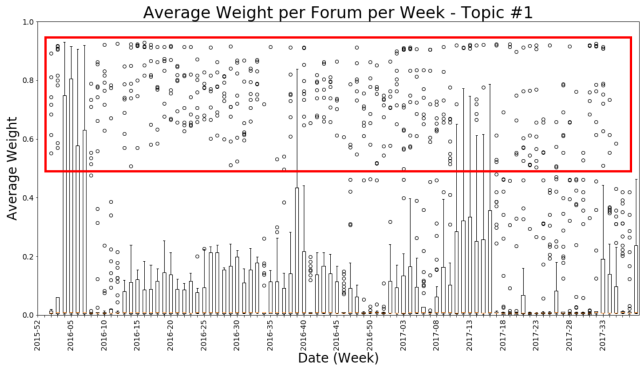


Figure 5: Boxplot Distribution of Proxy-server topic over the timespan of the corpus. Plotted are the 10 top forums in Cluster 2. Each point is the average weight of that topic for a forum for that week. The red box indicates the significant forums in each week.

We also looked at one of the most illicit drug topics, 11.Narcotics. We found that besides 11.Narcotics other prominent topics in the state where this topic is at its highest value are 9. Marijuana and 7.LSD. These are a few forums which exhibit this state: The Majestic Garden, the DarkNetReviews subreddit, and the HANSAdnmDarkNetMarket subreddit.

4.4.2 Proxy Servers. The second case study is around the discussion of proxy servers and the sale of services that can be used to game social media platforms, defraud ad networks, build botnets, and effectively launder illegal activity. We use the clustering from HMM to help identify which kinds of forums have high activity related to proxy servers. Compared to many of the other topics, topic 1.Locations is concerned primarily with the sale of proxy servers. When we examined cluster 2 from the HMM clustering, we noticed a significant activation of this topic. Interestingly, the ten most active forums in cluster 2 seem to capture most of the activation of the topic over time. As seen in Fig. 5, we examine the forum-weeks (points) with significant probability (above 0.50) for topic 1.Locations from forums in cluster 2 and recover approximately 11,000 documents which seem to be automated posts advertising the sale of access to proxies all over the world. These 11,000 documents come mostly from CSU and BlackWorld: both forums have subforums dedicated to the advertising of proxies and are in the top 10 active forums in the cluster. An excerpt from one of these posts is as follows: "...camarillo | ca | unknown | united states | checked at vn5socks.netlive...". When we look for documents pertaining to more specific uses of proxies, we find approximately 20 posts that directly mention "viewbot". Viewbotting is the act of using bots to artificially inflate the number of views on a social media profile (e.g. YouTube and Twitch). As it can be difficult to determine if a viewer is a human or a bot, this can potentially trick the social media platform into thinking a profile is more popular than it actually is and result in more attention than it would obtain organically. Alternate uses of viewbots are to watch video ads on a channel, artificially increasing ad revenue. An excerpt from one of these documents is as follows: "...i viewbotted my vid to 1k views and got 10 slaves..."

Another potential malicious use of proxy servers is for carding. Carding in this case refers to the fradulent use of other people's credit cards, personal and/or financial data to purchase goods, launder money, or generally steal an individual's money. Proxy servers are commonly used to "cash out" stolen credit card information by buying items like pre-paid gift cards through payment processors. In our corpus, there are a number of documents that mention carding. An example that suggests intent to use proxies for carding is as follows: "...proxies are often blacklisted when used for fraud so im looking for a source for fresh proxies for use for carding..."

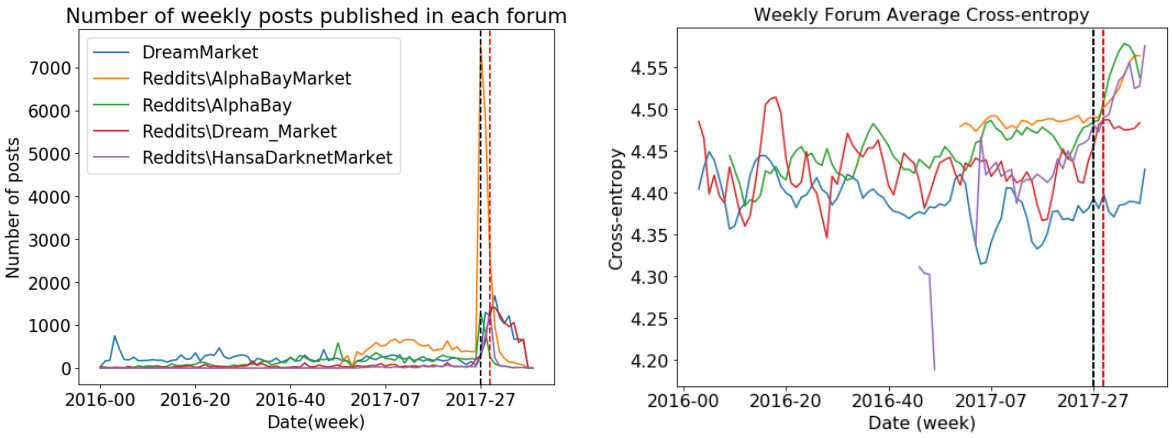
4.4.3 Marketplace Shutdowns. Our last case study regards the seizure of the AlphaBay marketplace by the FBI on July 4th 2017 and the seizure of the Hansa marketplace on July 20th 2017 by the Dutch NHTCU. We show cross-entropy, forum activity and transitions between states to analyze this case study.

Included in our d2web data are forums related to transactions and reviews of these marketplaces, including several private subreddits. We observed that a few of the forums in our dataset have peaks in activity around the same time. Figure 6(a) shows number of posts published in these forums. Forums related to AlphaBay and Hansa have peaks on the date of AlphaBay and Hansa closure respectively. Interestingly, a week after the Hansa closure Dream Market and the subreddits DreamMarket and DreamMarketDarknet had their highest value which suggests that users of these two big market places, AlphaBay and Hansa, have migrated to Dream Market. To check whether the forums respond to these events by changing the topics of discussion, we compute the cross-entropy. As seen in Figure 6(b) we see that two relevant forums, one about AlphaBay and one about DreamMarket, experience a change in their topic vectors after the shutdown (2017 week 27, or 07-04-2017). The cross-entropy increases around that time, suggesting growing difference from the aggregate topic distribution for their respective forums however changes in the volume of posts were more significant. We also observe state transitions in Dream Market forum in all three dates (AlphaBay and Hansa closure and the week after when there is a peak in activity for Dream Market related forums).

5 DISCUSSION & CONCLUSIONS

We used LDA to learn a rich set of latent components from a large corpus of documents spanning different topics and make use of a non-parametric HMM to better understand how those forums relate to one another in terms of the dynamics of their content. We then proceeded to use what we learn from both the states and the topics to identify specific posts discussing malicious activities that are understood to largely come from the dark and deep web.

This work can be extended in a number of different ways. First, one can use our framework to analyze new and unseen forums. Second, we can extend the framework to explicitly consider patterns of activity (e.g., frequency of posts) alongside the semantic information of the topic dynamics. Additionally, using average of the topic distributions of posts in a forum in one week to represent that forum will invariably smooth out events that have a small number of events associated with them. An alternate method that would be more sensitive to smaller fluctuations is to use the maximum value for a topic obtained from any post in that week.



(a) (b)

Figure 6: (a) Activity of forums relevant to AlphaBay and Hansa closure. (b) Smoothed weekly topic cross-entropy of forums. Cross-entropy is smoothed using a rolling average over 4 weeks. For both figures the black line indicates July 4th, 2017 when AlphaBay was seized. The red line indicates July 20th, 2017 when Hansa was seized.

A promising extension to our system would be to use a dynamic topic model that incorporates a birth-death process for the population of topics, allowing for new topics to emerge and for old topics to die off. Online communities move fast, and a topic model that includes data from even a couple months ago can become obsolete and hinder productive analysis.

Acknowledgments. This work was supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C- 0112, and the Defense Advanced Research Projects Agency (DARPA), contract number W911NF-17-C-0094. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, DARPA, or the U.S. Government.

REFERENCES

[1] Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian. 2015. BiSAL–A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security. *Digital Investigation* 14 (2015), 53–62.

[2] Mohammed Almukaynizi, Eric Nunes, Krishna Dharaiya, Manoj Senguttuvan, Jana Shakarian, and Paulo Shakarian. 2017. Proactive identification of exploits in the wild through vulnerability mentions online. In *Cyber Conflict (CyCon US), 2017 International Conference on*. IEEE, 82–88.

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

[4] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 113–120.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[6] Hsinchun Chen. 2008. Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In *Intelligence and Security Informatics*.

[7] Hsinchun Chen, Wingyan Chung, Jialun Qin, Edna Reid, Marc Sageman, and Gabriel Weimann. 2008. Uncovering the dark Web: A case study of Jihad on the Web. *JASIST* 59, 8 (2008), 1347–1359.

[8] Ashok Deb, Kristina Lerman, and Emilio Ferrara. 2018. Predicting Cyber-Events by Leveraging Hacker Sentiment. *Information* 9, 11 (2018).

[9] Emily Fox, Michael I Jordan, Erik B Sudderth, and Alan S Willsky. 2009. Sharing features among dynamical systems with beta processes. In *NIPS*. 549–557.

[10] Emily B Fox, Michael C Hughes, Erik B Sudderth, Michael I Jordan, et al. 2014. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics* 8, 3 (2014).

[11] Robert W Gehl. 2016. Power/freedom on the dark web: A digital ethnography of the Dark Web Social Network. *new media & society* 18, 7 (2016), 1219–1235.

[12] Zoubin Ghahramani and Thomas L Griffiths. 2006. Infinite latent feature models and the Indian buffet process. In *NIPS*. 475–482.

[13] Palash Goyal, KSM Hossain, Ashok Deb, Nazgol Tavabi, Nathan Bartley, Andr'es Abeliuk, Emilio Ferrara, and Kristina Lerman. 2018. Discovering Signals from Web Sources to Predict Cyber Attacks. *arXiv preprint arXiv:1806.03342* (2018).

[14] Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. 2005. Integrating topics and syntax. In *NIPS*. 537–544.

[15] Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Artificial intelligence and statistics*. 163–170.

[16] Nils Lid Hjort et al. 1990. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* 18, 3 (1990), 1259–1294.

[17] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).

[18] George Hurlburt. 2017. Shining Light on the Dark Web. *IEEE Computer* 50, 4 (2017), 100–105.

[19] Gastón L'huillier, Hector Alvarez, Sebastián A Ríos, and Felipe Aguilera. 2011. Topic-based social network analysis for virtual communities of interests in the dark web. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 66–73.

[20] Erik Linstead, Cristina Lopes, and Pierre Baldi. 2008. An application of latent Dirichlet allocation to analyzing software evolution. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*. IEEE, 813–818.

[21] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[22] Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart, and Paulo Shakarian. 2016. Darknet and deepnet mining for proactive cybersecurity threat intelligence. *arXiv preprint arXiv:1607.08583* (2016).

[23] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Workshop on New Challenges for NLP Frameworks*. 45–50.

[24] Sebastián A Ríos and Ricardo Muñoz. 2012. Dark Web portal overlapping community detection based on topic models. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*.

[25] John Robertson, Ahmad Diab, Ericsson Marin, Eric Nunes, Vivin Paliath, Jana Shakarian, and Paulo Shakarian. 2017. *Darkweb Cyber Threat Intelligence Mining*.

[26] Jana Shakarian, Andrew T Gunn, and Paulo Shakarian. 2016. Exploring malicious hacker forums. In *Cyber Deception*. Springer, 259–282.

[27] Kyle Soska and Nicolas Christin. 2015. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. In *USENIX Security Symposium*.

[28] Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. DarkEmbed: Exploit Prediction With Neural Language Models. In *AAAI*.

[29] Romain Thibaux and Michael I Jordan. 2007. Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*. 564–571.

[30] Seshadri Tirunillai and Gerard J Tellis. 2014. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research* 51, 4 (2014), 463–479.

[31] Jennifer Xu and Hsinchun Chen. 2008. The topology of dark networks. *Commun. ACM* 51, 10 (2008), 58–65.

Bot Electioneering Volume: Visualizing Social Bot Activity During Elections

Kai-Cheng Yang
Center for Complex Networks and Systems Research, Indiana University
Bloomington, Indiana
yangkc@iu.edu

Pik-Mai Hui
Center for Complex Networks and Systems Research, Indiana University
Bloomington, Indiana
huip@iu.edu

Filippo Menczer
Center for Complex Networks and Systems Research, Indiana University
Bloomington, Indiana
cnets.indiana.edu/fil

ABSTRACT

It has been widely recognized that automated bots may have a significant impact on the outcomes of national events. It is important to raise public awareness about the threat of bots on social media during these important events, such as the 2018 US midterm election. To this end, we deployed a web application to help the public explore the activities of likely bots on Twitter on a daily basis. The application, called Bot Electioneering Volume (BEV), reports on the level of likely bot activities and visualizes the topics targeted by them. With this paper we release our code base for the BEV framework, with the goal of facilitating future efforts to combat malicious bots on social media.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

bot activity; Twitter; elections

ACM Reference Format:
Kai-Cheng Yang, Pik-Mai Hui, and Filippo Menczer. 2019. Bot Electioneering Volume: Visualizing Social Bot Activity During Elections. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316499

1 INTRODUCTION

Social bots have drawn great attention from the public recently. They are accounts on social media platforms controlled at least in part by algorithms to generate/share/retweet content and interact with human users [6]. The automated nature of social bots makes it easy to achieve scalability, with which a single person is capable to control thousands of accounts on one or more social media platforms. When needed, these social bots can work collectively to manipulate the public by promoting certain accounts or opinions.

Being social animals, human users are inevitably vulnerable to the efforts of the social bots. Studies have shown ubiquitous social bots [13] distort online discussions, and particularly those about politics. During the 2010 US midterm election, primitive social bots were used to attack some candidates [7] and spread tweets with links to fake news websites [8]. A similar pattern emerged in the 2016 US presidential election, only with more sophisticated bots that aimed to effectively push their messages to the target audience [1]. In particular, bots were most active in the core of the misinformation-sharing network [11] and effectively amplified the spread of low-credibility content by posting it within seconds and by targeting influential accounts [10]. Analogous automated campaigns were reported in countries around the globe [5, 12].

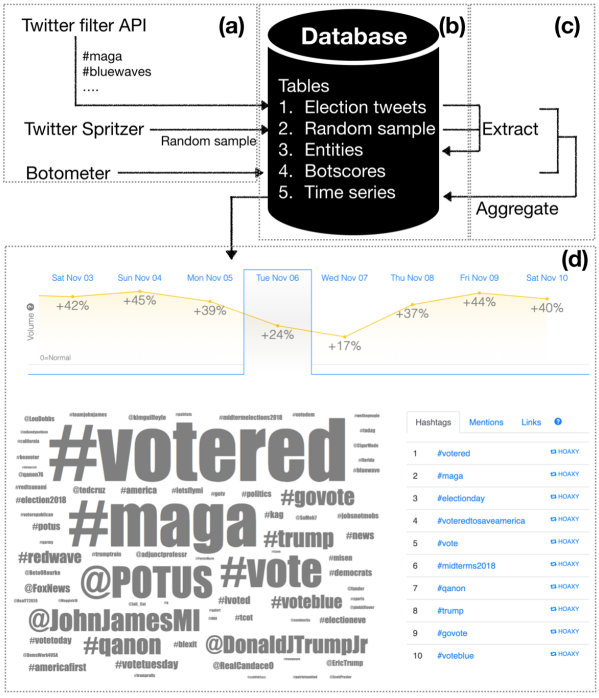


Figure 1: Illustration of BEV's (a) crawler, (b) database, (c) analyzer, and screen shot of (d) front-end interface. The upper panel of the frontend shows the Bot Electioneering Volume for the past 8 days. Users can select a day of interest to explore the top topics of that day. The bottom panels show a tag cloud and entity lists for the selected day. The tag cloud presents entities all together, with the size of each entity proportional to how often it is tweeted by bots. The entity lists display hashtags, mentions, and links ranked by how often each is tweeted by bots.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316499

Here we present Bot Electioneering Volume (BEV), a platform that visualizes the volume generated by bots and the corresponding targeted topics. BEV tracks online traffic centered around elections from Twitter by feeding the streaming API with a list of selected hashtags. By incorporating the bot-detection ability of Botometer [4, 13, 14], BEV is able to distinguish between content generated by likely bots and humans. The measurement of average bot activity is then compared with random samples of tweets to produce a number that quantifies electioneering activity by bots. BEV also collects content topics, including hashtags, mentions, and URLs shared by likely bots, and reports on their relative volumes.

BEV (botometer.iuni.iu.edu/bev) monitored public tweets about the 2018 US midterm elections between October 22, 2018 and Dec 30, 2018. During the collection period, BEV drew over 3,000 visits. An archive of the data from Oct to Dec 2018 remains publicly available for retrospective inspections. We plan to activate BEV again for future elections. Our goal is to raise public awareness of bot activities and their impact during elections in the past, and more importantly those in the future.

2 SYSTEM DESIGN

The BEV system contains 4 major parts: a crawler, a database, an analyzer, and a front-end interface, as illustrated in Figure 1.

The crawler is in charge of tracking Twitter's filtering API for public election-related tweets, querying Twitter's Spritzer API for random samples of public tweets, and fetching bot scores. Crawled data is stored in the database. The analyzer then extracts the required information and generates the statistics for the visualization at the application frontend. The frontend has three major parts: the Bot Electioneering Volume timeline, a tag cloud, and entity lists. The Bot Electioneering Volume measures the activity of likely bots, while the tag could and entity lists display the topics that are most tweeted by likely bots. By clicking on the links in an entity list, users are directed to Hoaxy (hoaxy.iuni.iu.edu) [9], where they can explore more in-depth visualizations of the influence of bots around the entities on Twitter in the recent minutes/hours/days.

The data collection runs in a streaming fashion, but fetching bot scores and analyzing the data take time. The front-end interface is updated every 4 hours to reflect newly incoming data.

2.1 Data collection

The collection of election-related tweets is crucial to our application. Our collection process starts with a set of election-related hashtags that are tracked using the Twitter filtering API. We seeded our set of hashtags with several widely-used political hashtags including #2018midterms, #maga, and #bluewave. We then repeatedly expanded the set with co-occuring hashtags [2, 3], resulting in a set of 110 hashtags. From this set we manually removed 6 hashtags that are general and irrelevant to election. We also added the hashtags for each US state's Senate race: #casen, #nysen, and so on. The full list of hashtags can be found in the FAQ page of the BEV website. This methodology allows our system to collect most tweets with newly emerging election hashtags, because it is likely that these tweets contain some of the hashtags in our list as well.

Twitter's free filtering API offers at most 1% of Twitter's traffic. We estimated that the traffic captured by our method is about 0.3–0.5% of Twitter's complete traffic (see Figure 2(a)). This means that instead of a sample, BEV collects and visualizes bot activities based on *all* of the targeted election-related tweets.

2.2 Bot identification

BEV uses Botometer (botometer.iuni.iu.edu) [4, 13, 14] to obtain bot scores for Twitter accounts involved in election-related discourse. Botometer is a supervised machine learning algorithm that considers more than a thousand features about an account and its activity to estimate the likelihood that the account is automated. The newest version of Botometer incorporates various training datasets and has accuracy of 0.97 AUC under 5-fold cross-validation [14]. We consider accounts with bot score above 4 (on a 5-point scale) as social bots. This is a fairly conservative threshold choice, corresponding to a posterior probability of automation near 50% [14] based on a 15% prior probability [13].

2.3 Bot activity measurement

To measure the bot electioneering volume, we first take daily averages of the bot scores of accounts generating political tweets, weighted by their tweet frequencies. Considering that spamming the same message is a common strategy for bots, the weighted average better highlights the amount of bot-generated content and their potential influence. To obtain a baseline of bot activity, we produce the same weighted average for random tweets from Twitter's Spritzer API, with a rate of 1,000 tweets per hour. The daily average bot scores for election and random tweets are shown in Figure 2(b).

The Bot Electioneering Volume is defined by the relative difference between the two averages:

$$\mathrm{BEV} = \frac{S_{\mathrm{Electoral}} - S_{\mathrm{RandomSample}}}{S_{\mathrm{RandomSample}}}, \tag{1}$$

where $S_{\mathrm{Electoral}}$ and $S_{\mathrm{RandomSample}}$ represent the average bot scores of electoral and random tweets, respectively. BEV is shown as a percentage difference in the front-end interface, as shown in Figure 1(d). Figure 2(c) plots the Bot Electioneering Volume of the whole midterm elections period.

In the design stage of BEV, we considered different metrics by replacing the average bot score S in Eq. 1 with the median score and the proportion of tweets by bots. The Bot Electioneering Volume based on median values of the scores yields patterns similar to the average score version. However, the version based on the proportion of tweets by bots (denoted by BEV2) shows different trends. As shown in Figure 3(a), the baseline proportion of tweets by bots decreases steadily after the midterm elections. Consequently, the BEV2 value increases (Figure 3(b)). Since the proportion of tweets by likely bots in the random sample is not stable, we decided to deploy the BEV metric based on average bot scores.

The BEV front-end interface also visualizes the topics targeted by likely bots. For simplicity, we only extract hashtags, mentioned accounts, and links from the tweets generated by bots. The targeted topics are represented in a tag cloud and entity lists.

Figure 2: (a) Number of election tweets and unique users for each day. (b) Average bot scores for election tweets and random sample. (c) BEV timeline. The 2018 US midterm elections day, Nov 6, is highlighted by a vertical dashed line.

3 DISCUSSION

BEV reveals many interesting patterns. In terms of the discussion intensity, the number of election tweets before the election was about twice as many as the number after the election day peak, as shown in Figure 2(a). For bot activity, all the different metrics — average bot score, median bot score, and proportion of tweets by likely bots — are much higher for election tweets, suggesting that bots are actively generating election-related content; they are indeed employed for electioneering.

The average bot scores of election-related tweets fluctuated from day to day. The BEV on November 6 and 7 was drastically lower than on other days. Considering the spike in election tweets, we hypothesize that the bot activity was diluted by the huge amount of normal users around election day. Furthermore, the average bot scores of election-related tweets after the election are generally higher than those before the election. This is perhaps because many human users ceased to tweet political content after the election, but the bots kept working. BEV cannot reflect the changes in total volume of tweets because it is based on averaging the daily scores. For the future, a measurement that can leverage both the average bot score and the volume of tweets may be preferable to better represent the volume generated by bots.

4 CONCLUSION

We offer a real-time tool to visualize the electioneering activities of social bots. Open-source code for BEV is at github.com/IUNetSci/BEV.

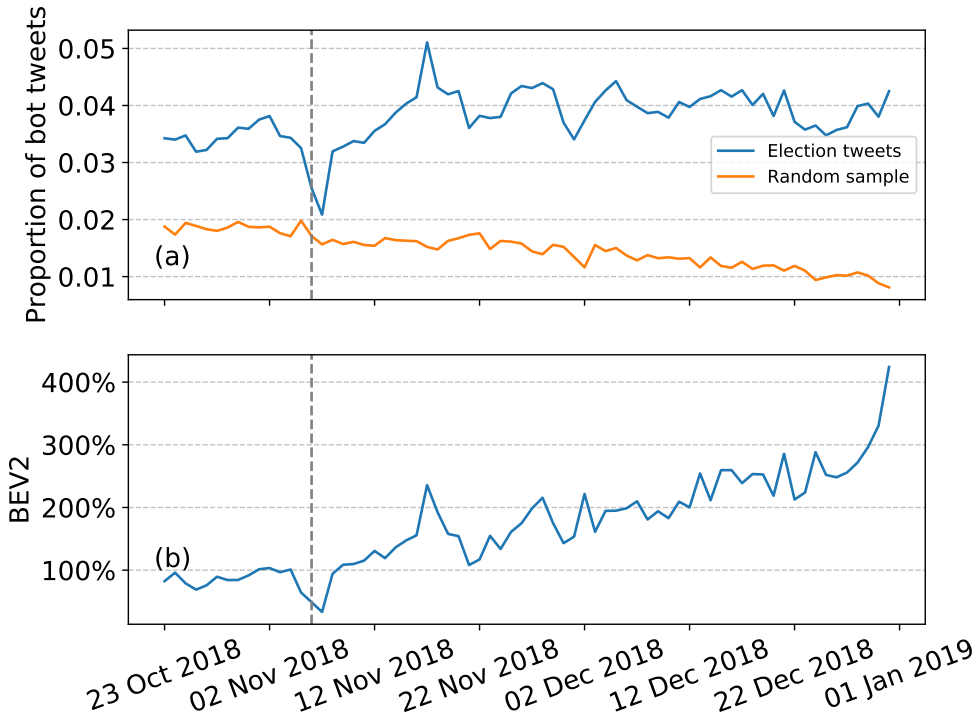


Figure 3: (a) Proportion of tweets by bots for election tweets and random sample. (b) BEV2.

Around the 2018 US midterm elections, bot electioneering was rampant. The great majority of content amplified by likely bots was on the conservative side of the political spectrum. It remains to be seen if this will change in the future.

The tool enables the public to gain a sense of the organic nature of the online discussion regarding elections and spot possibly polluted content. By performing the hashtag list generation procedure with different seeds, BEV can be adapted to target future US elections, as well as elections and events in other countries.

Acknowledgments. We are grateful to the IU Network Science Institute (iuni.iu.edu) and the AWS Cloud Credits for Research program for supporting the BEV infrastructure and to Twitter for their free APIs. Botometer was developed with support from the Democracy Fund.

REFERENCES

[1] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).

[2] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring #gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th international conference on world wide web companion*. International World Wide Web Conferences Steering Committee, IW3C2, Perth, 1285–1290.

[3] Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2012. Partisan asymmetries in online political activity. *EPJ Data Science* 1 (2012), 6. https://doi.org/10.1140/epjds6

[4] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A system to evaluate social bots. In *Proc. 25th Intl. Conf. Companion on World Wide Web*. IW3CSC, Montreal, 273–274.

[5] Emilio Ferrara. 2017. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday* 22, 8 (2017).

[6] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[7] Panagiotis T Metaxas and Eni Mustafaraj. 2012. Social media and the elections. *Science* 338, 6106 (2012), 472–473.

[8] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. In *Proc. Intl. Conf. on Weblogs and Social Media (ICWSM)*. AAAI, Barcelona, 297–304.

[9] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proc. 25th Intl. Conf. Companion on World Wide Web*. IW3CSC, Montreal, 745–750.

[10] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nat. Commun.* 9 (2018), 4787. https://doi.org/10.1038/s41467-018-06930-7

[11] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2018. Anatomy of an online misinformation network. *PLoS ONE* 13, 4 (2018), e0196087. https://doi.org/10.1371/journal.pone.0196087

[12] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. U.S.A.* 115, 49 (2018), 12435–12440.

[13] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proc. Intl. Conf. on Web and Social Media (ICWSM)*. AAAI, Montreal, 280–289.

[14] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with AI to counter social bots. *Human Behavior and Emerging Technologies* 1, 1 (2019). https://doi.org/10.1002/hbe2.115

Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web

Savvas Zannettou
Cyprus University of Technology
sa.zannettou@edu.cut.ac.cy

Tristan Caulfield
University College London
t.caulfield@ucl.ac.uk

Emiliano De Cristofaro
University College London
e.decristofaro@ucl.ac.uk

Michael Sirivianos
Cyprus University of Technology
michael.sirivianos@cut.ac.cy

Gianluca Stringhini
Boston University
gian@bu.edu

Jeremy Blackburn
University of Alabama at Birmingham
blackburn@uab.edu

ABSTRACT

Over the past couple of years, anecdotal evidence has emerged linking coordinated campaigns by state-sponsored actors with efforts to manipulate public opinion on the Web, often around major political events, through dedicated accounts, or "trolls." Although they are often involved in spreading disinformation on social media, there is little understanding of how these trolls operate, what type of content they disseminate, and most importantly their influence on the information ecosystem.

In this paper, we shed light on these questions by analyzing 27K tweets posted by 1K Twitter users identified as having ties with Russia's Internet Research Agency and thus likely state-sponsored trolls. We compare their behavior to a random set of Twitter users, finding interesting differences in terms of the content they disseminate, the evolution of their account, as well as their general behavior and use of Twitter. Then, using Hawkes Processes, we quantify the influence that trolls had on the dissemination of news on social platforms like Twitter, Reddit, and 4chan. Overall, our findings indicate that Russian trolls managed to stay active for long periods of time and to reach a substantial number of Twitter users with their tweets. When looking at their ability of spreading news content and making it viral, however, we find that their effect on social platforms was minor, with the significant exception of news published by the Russian state-sponsored news outlet RT (Russia Today).

CCS CONCEPTS

• **General and reference → General conference proceedings**; **Measurement**; • **Mathematics of computing → Probabilistic algorithms**; • **Networks → Social media networks**; **Online social networks**.

KEYWORDS

disinformation, trolls, social networks, twitter, reddit, 4chan

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316495

ACM Reference Format:
Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316495

1 INTRODUCTION

Recent political events and elections have been increasingly accompanied by reports of disinformation campaigns attributed to state-sponsored actors [7]. In particular, "troll farms," allegedly employed by Russian state agencies, have been actively commenting and posting content on social media to further the Kremlin's political agenda [23]. In late 2017, the US Congress started an investigation on Russian interference in the 2016 US Presidential Election, releasing the IDs of 2.7K Twitter accounts identified as Russian trolls.

Despite the growing relevance of state-sponsored disinformation, the activity of accounts linked to such efforts has not been thoroughly studied. Previous work has mostly looked at campaigns run by bots [7, 10, 20]; however, automated content diffusion is only a part of the issue, and in fact recent research has shown that human actors are actually key in spreading false information on Twitter [21]. Overall, many aspects of state-sponsored disinformation remain unclear, e.g., how do state-sponsored trolls operate? What kind of content do they disseminate? And, perhaps more importantly, is it possible to quantify the influence they have on the overall information ecosystem on the Web?

In this paper, we aim to address these questions, by relying on the set of 2.7K accounts released by the US Congress as ground truth for Russian state-sponsored trolls. From a dataset containing all tweets released by the 1% Twitter Streaming API, we search and retrieve 27K tweets posted by 1K Russian trolls between January 2016 and September 2017. We characterize their activity by comparing to a random sample of Twitter users. Then, we quantify the influence of these trolls on the greater Web, looking at occurrences of URLs posted by them on Twitter, 4chan [11], and Reddit, which we choose since they are impactful actors of the information ecosystem [29]. Finally, we use Hawkes Processes [15] to model the influence of each Web community (i.e., Russian trolls on Twitter, overall Twitter, Reddit, and 4chan) on each other.

Main findings. Our study leads to several key observations:

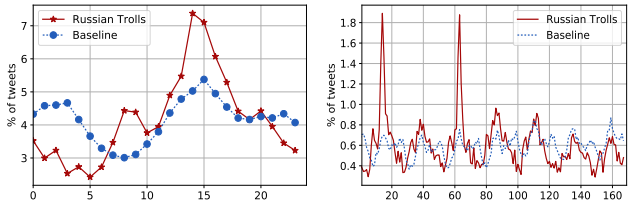
(1) Trolls actually bear very small influence in making news go viral on Twitter and other social platforms alike. A noteworthy exception are links to news originating from RT (Russia Today), a state-funded news outlet: indeed, Russian trolls are quite effective in "pushing" these URLs on Twitter and other social networks.
(2) The main topics discussed by Russian trolls target very specific world events (e.g., Charlottesville protests) and organizations (such as ISIS), and political threads related to Donald Trump and Hillary Clinton.
(3) Trolls adopt different identities over time, i.e., they "reset" their profile by deleting their previous tweets and changing their screen name/information.
(4) Trolls exhibit significantly different behaviors compared to other (random) Twitter accounts. For instance, the locations they report concentrate in a few countries like the USA, Germany, and Russia, perhaps in an attempt to appear "local" and more effectively manipulate opinions of users from those countries. Also, while random Twitter users mainly tweet from mobile versions of the platform, the majority of the Russian trolls do so via the Web Client.

2 DATASETS

Russian trolls. We start from the 2.7K Twitter accounts suspended by Twitter because of connections to Russia's Internet Research Agency. The list of these accounts was released by the US Congress as part of their investigation of the alleged Russian interference in the 2016 US presidential election, and includes both Twitter's *user ID* (which is a numeric unique identifier associated to the account) and the *screen name*.[1] From a dataset storing all tweets released by the 1% Twitter Streaming API, we search for tweets posted between January 2016 and September 2017 by the user IDs of the trolls. Overall, we obtain 27K tweets from 1K out of the 2.7K Russian trolls. Note that the criteria used by Twitter to identify these troll accounts are not public. What we do know is that this is not the complete set of active Russian trolls, because 6 days prior to this writing Twitter announced they have discovered over 1K more troll accounts.[2] Nonetheless, it constitutes an invaluable "ground truth" dataset enabling efforts to shed light on the behavior of state-sponsored troll accounts.

Baseline dataset. We also compile a list of random Twitter users, while ensuring that the distribution of the average number of tweets per day posted by the random users is similar to the one by trolls. To calculate the average number of tweets posted by an account, we find the first tweet posted after January 1, 2016 and retrieve the overall tweet count. This number is then divided by the number of days since account creation. Having selected a set of 1K random users, we then collect all their tweets between January 2016 and September 2017, obtaining a total of 96K tweets. We follow this approach as it gives a good approximation of posting behavior, even though it might not be perfect, since (1) Twitter accounts can become more or less active over time, and (2) our datasets are based on the 1% Streaming API, thus, we are unable to control the number of tweets we obtain for each account.

[1] See https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf
[2] https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html



(a) Hour of Day (b) Hour of Week

Figure 1: Temporal characteristics of tweets from Russian trolls and random Twitter users.

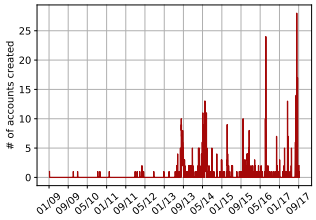


Figure 2: Number of Russian troll accounts created per day.

3 ANALYSIS

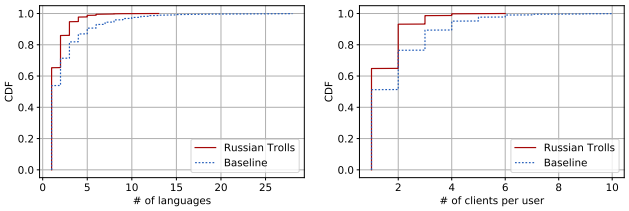
Temporal analysis. We observe that Russian trolls are continuously active on Twitter between January, 2016 and September, 2017, with a peak of activity just before the second US presidential debate (October 9, 2016). Fig. 1(a) shows that most tweets from the trolls are posted between 14:00 and 15:00 UTC. In Fig. 1(b), we also report temporal characteristics based on hour of the week, finding that both datasets follow a diurnal pattern, while trolls' activity peaks around 14:00 and 15:00 UTC on Mondays and Wednesdays. Considering that Moscow is three hours ahead UTC, this distribution does not rule out that tweets might actually be posted from Russia.

Account creation. Next, we examine the dates when the trolls infiltrated Twitter, by looking at the account creation dates. From Fig. 2, we observe that 71% of them are actually created before 2016. There are some interesting peaks, during 2016 and 2017: for instance, 24 accounts are created on July 12, 2016, approx. a week before the Republican National Convention (when Donald Trump received the nomination), while 28 appear on August 8, 2017, a few days before the infamous Unite the Right rally in Charlottesville. Taken together, this might be evidence of coordinated activities aimed at manipulating users' opinions with respect to specific events.

Account characteristics. We also shed light on the troll account profile information. In Table 1, we report the top ten words appearing in the screen names and the descriptions of Russian trolls, as well as character 4-grams for screen names and word bigrams for profile descriptions. Interestingly, a substantial number of Russian trolls pose as news outlets, evident from the use of the term "news" in both the screen name (1.3%) and the description (10.7%). Also, it seems they attempt to increase the number of their followers, thus their reach of Twitter users, by nudging users to follow them (see, e.g., "follow me" appearing in almost 8% of the accounts). Finally,

Table 1: Top 10 words found in Russian troll screen names and account descriptions. We also report character 4-grams for the screen names and word bigrams for the description.

Screen Name				Description			
Word	(%)	4-gram	(%)	Word	(%)	Word bigram	(%)
news	1.3%	news	1.5%	news	10.7%	follow me	7.8%
bote	1.2%	line	1.5%	follow	10.7%	breaking news	2.6%
online	1.1%	blac	1.3%	conservative	8.1%	news aus	2.1%
daily	0.8%	bote	1.3%	trump	7.8%	uns in	2.1%
today	0.6%	rist	1.1%	und	6.2%	deiner stdt	2.1%
ezekiel2517	0.6%	nlin	1.1%	maga	5.9%	die news	2.1%
maria	0.5%	onli	1.0%	love	5.8%	wichtige und	2.1%
black	0.5%	lack	1.0%	us	5.3%	nachrichten aus	2.1%
voice	0.4%	bert	1.0%	die	5.0%	aus deiner	2.1%
martin	0.4%	poli	1.0%	nachrichten	4.3%	die dn	2.1%



(a) (b)

Figure 3: CDF of number of (a) languages used (b) clients used per user.

Table 2: Top 10 Twitter clients (as % of tweets).

Client (Trolls)	(%)	Client (Baseline)	(%)
Twitter Web Client	50.1%	TweetDeck	32.6%
twitterfeed	13.4%	Twitter for iPhone	26.2%
Twibble.io	9.0%	Twitter for Android	22.6%
IFTTT	8.6%	Twitter Web Client	6.1%
TweetDeck	8.3%	GrabInbox	2.0%
NovaPress	4.6%	Twitter for iPad	1.4%
dlvr.it	2.3%	IFTTT	1.0%
Twitter for iPhone	0.8%	twittbot.net	0.9%
Zapier.com	0.6%	Twitter for BlackBerry	0.6%
Twitter for Android	0.6%	Mobile Web (M2)	0.4%

10.3% of the Russian trolls describe themselves as Trump supporters: "trump" and "maga" (Make America Great Again, one of Trump campaign's main slogans).

Language. Looking at the language (as provided via the Twitter API) of the tweets posted by Russian trolls, we find that most of them (61%) are in English, although a substantial portion are in Russian (27%), and to a lesser extent in German (3.5%). In Fig. 3(a), we plot the cumulative distribution function (CDF) of the number of different languages for each user: 64% of the Russian trolls post all their tweets in only one language, compared to only 54% for random users. Overall, by comparing the two distributions, we observe that random Twitter users tend to use more languages in their tweets compared to the trolls.

Client. Finally, we analyze the clients used to post tweets. We do so since previous work [5] shows that the client used by official or professional accounts are quite different that the ones used by regular users. Table 2 reports the top 10 clients for both Russian trolls and baseline users. We find the latter prefer to use Twitter

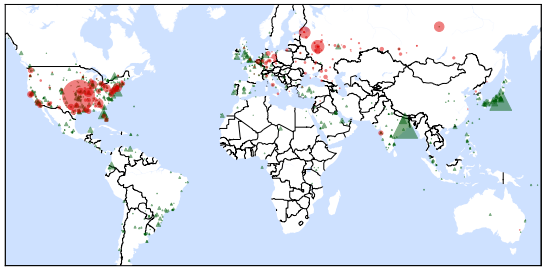


Figure 4: Distribution of reported locations for tweets by Russian trolls (red circles) and baseline (green triangles).

clients for mobile devices (48%) and the TweetDeck dashboard (32%), whereas, the former mainly use the Web client (50%). We also assess how many different clients Russian trolls use throughout our dataset: in Fig. 3(b), we plot the CDF of the number of clients used per user. We find that 65% of the Russian trolls use only one client, 28% of them two different clients, and the rest more than three, which is overall less than the random baseline users.

Location. We then study users' location, relying on the self-reported location field in their profiles. Note that users not only may leave it empty, but also change it any time they like, so we look at locations for each tweet. We retrieve it for 75% of the tweets by Russian trolls, gathering 261 different entries, which we convert to a physical location using the Google Maps Geocoding API. In the end, we obtain 178 unique locations for the trolls, as depicted in Fig. 4 (red circles). The size of the circles on the map indicates the number of tweets that appear at each location. We do the same for the baseline, getting 2,037 different entries, converted by the API to 894 unique locations. We observe that most of the tweets from Russian trolls come from locations within the USA and Russia, and some from European countries, like Germany, Belgium, and Italy. On the other hand, tweets in our baseline are more uniformly distributed across the globe, with many tweets from North and South America, Europe, and Asia. This suggests that Russian trolls may be pretending to be from certain countries, e.g., USA or Germany, aiming to pose as locals and better manipulate opinions. This explanation becomes more plausible when we consider that a plurality of trolls' tweets have their location set as a generic form of "US," as opposed to a specific city, state, or even region. Interestingly, the 2nd, 3rd, and 4th most popular location for trolls to tweet from are Moscow, St. Petersburg, and a generic form of "Russia." We also assess whether users change their country of origin based on the self-reported location: only a negligible percentage (1%) of trolls change their country, while for the baseline the percentage is 16%.

Media. We then assess whether Russian trolls use images and videos in a different way than random baseline users. For Russian trolls (resp., baseline accounts), 66% (resp., 73%) of the tweets include no images, 32% (resp., 18%) exactly one image, and 2% (resp., 9%) more than one. This suggests that Russian trolls disseminate a considerable amount of information via single-image tweets. As for videos, only 1.5% of the tweets from Russian trolls includes a video, as opposed to 6.4% for baseline accounts.

Table 3: Top 20 hashtags in tweets from Russian trolls and baseline users.

Trolls				Baseline			
Hashtag	(%)	Hashtag	(%)	Hashtag	(%)	Hashtag	(%)
news	7.2%	US	0.7%	iHeartAwards	1.8%	UrbanAttires	0.6%
politics	2.6%	tcot	0.6%	BestFanArmy	1.6%	Vacature	0.6%
sports	2.1%	PJNET	0.6%	Harmonizers	1.0%	mPlusPlaces	0.6%
business	1.4%	entertainment	0.5%	iOSApp	0.9%	job	0.5%
money	1.3%	top	0.5%	JouwBaan	0.9%	Directioners	0.5%
world	1.2%	topNews	0.5%	vacature	0.9%	JIMIN	0.5%
MAGA	0.8%	ISIS	0.4%	KCA	0.9%	PRODUCE101	0.5%
health	0.8%	Merkelmussbleiben	0.4%	Psychic	0.8%	VoteMainFPP	0.5%
local	0.7%	IslamKills	0.4%	RT	0.8%	Werk	0.4%
BlackLivesMatter	0.7%	breaking	0.4%	Libertad2016	0.6%	dts	0.4%

Hashtags. Our next step is to study the use of hashtags in tweets. Russian trolls use at least one hashtag in 32% of their tweets, compared to 10% for the baseline. Overall, we find 4.3K and 7.1K unique hashtags for trolls and random users, respectively, with 74% and 78% of them only appearing once. In Table 3, we report the top 20 hashtags for both datasets. Trolls appear to use hashtags to disseminate news (7.2%) and politics (2.6%) related content, but also use several that might be indicators of propaganda and/or controversial topics, e.g., #ISIS, #IslamKills, and #BlackLivesMatter. For instance, we find some notable examples including: "We just have to close the borders, 'refugees' are simple terrorists #IslamKills" on March 22, 2016, "#SyrianRefugees ARE TERRORISTS from #ISIS #IslamKills" on March 22, 2016, and "WATCH: Here is a typical #BlackLivesMatter protester: 'I hope I kill all white babes!' #BatonRouge <url>" on July 17, 2016.

We also study when these hashtags are used by the trolls, finding that most of them are well distributed over time. However, there are some interesting exceptions, e.g., with #Merkelmussbleiben (a hashtag seemingly supporting Angela Merkel) and #IslamKills. Specifically, tweets with the former appear exclusively on July 21, 2016, while the latter on March 22, 2016, when a terrorist attack took place at Brussels airport. These two examples illustrate how the trolls may be coordinating to push specific narratives on Twitter.

Mentions. We find that 46% of trolls' tweets include *mentions* to 8.5K unique Twitter users. This percentage is much higher for the random baseline users (80%, to 41K users). Table 4 reports the 20 top mentions we find in tweets from Russian trolls and baseline users. We find several Russian accounts, like 'leprasorium' (a popular Russian account that mainly posts memes) in 2% of the mentions, as well as popular politicians like 'realDonaldTrump' (0.6%). The practice of mentioning politicians on Twitter may reflect an underlying strategy to mutate users' opinions regarding a particular political topic, which has been also studied in previous work [2].

URLs. We then analyze the URLs included in the tweets. First of all, we note that 53% of the trolls' tweets include at least a URL, compared to only 27% for the random baseline. There is an extensive presence of URL shorteners for both datasets, e.g., bit.ly (12% for trolls and 26% for the baseline) and ift.tt (10% for trolls and 2% for the baseline), therefore, in November 2017, we visit each URL to obtain the final URL after all redirections. In Fig. 5, we plot the CDF of the number of URLs per unique domain. We observe that Russian trolls disseminate more URLs in their tweets compared to the baseline. This might indicate that Russian trolls include URLs to increase their credibility and positive user perception; indeed, [9] show that adding a URL in a tweet correlates with higher credibility scores. Also, in Table 6, we report the top 20 domains for both Russian trolls and the baseline. Most URLs point to content within Twitter itself; 13% and 35%, respectively. Links to a number of popular social networks like YouTube (1.8% and 4.2%, respectively) and Instagram (1.5% and 1.9%) appear in both datasets. We also note that among the top 20 domains, there are also a number of news outlets linked from trolls' tweets, e.g., Washington Post (0.7%), Seattle Times (0.7%), and state-sponsored news outlets like RT (0.8%) in trolls' tweets, but much less so from the baseline.

Table 4: Top 20 mentions in tweets from trolls and baseline.

Trolls				Baseline			
Mention	(%)	Mention	(%)	Mention	(%)	Mention	(%)
leprasorium	2.1%	postsovet	0.4%	TasbihIstighfar	0.3%	RasSpotlights	0.1%
zubovnik	0.8%	DLGreez	0.4%	raspotlights	0.2%	GenderReveals	0.1%
realDonaldTrump	0.6%	DanaGeezus	0.4%	FunnyBrawls	0.2%	TattedChanel	0.1%
midnight	0.6%	ruopentwit	0.3%	YouTube	0.2%	gemvius	0.1%
blicqer	0.6%	Spoontamer	0.3%	Harry_Styles	0.2%	DrizzyNYC__	0.1%
gloed_up	0.6%	YouTube	0.3%	shortdancevids	0.2%	August_Alsina_	0.1%
wylsacom	0.5%	ChrixMorgan	0.3%	UrbanAttires	0.2%	RihannaBITCH_	0.1%
TalibKweli	0.4%	sergeylazarev	0.3%	BTS_twt	0.2%	sexualfeed	0.1%
zvezdanews	0.4%	RT_com	0.3%	KylieJenner_NYC	0.2%	PetsEvery30	0.1%
GiselleEvns	0.4%	kozheed	0.3%	BaddiessNation	0.2%	IGGYAZALEAoO	0.1%

Table 5: Top 10 domains in tweets from trolls and the baseline.

Domain (Trolls)	(%)	Domain (Baseline)	(%)
twitter.com	12.81%	twitter.com	35.51%
reportsecret.com	7.02%	youtube.com	4.21%
riafan.ru	3.42%	vine.co	3.94%
politexpert.net	2.10%	factissues.com	3.24%
youtube.com	1.88%	blogspot.com.cy	1.92%
vk.com	1.58%	instagram.com	1.90%
instagram.com	1.53%	facebook.com	1.68%
yandex.ru	1.50%	worldstarr.info	1.47%
infreactor.org	1.36%	trendytopic.info	1.39%
cbslocal.com	1.35%	minibird.jp	1.25%

Table 6: Top 20 domains included in tweets from Russian trolls and baselines users.

Domain (Trolls)	(%)	Domain (Baseline)	(%)
twitter.com	12.81%	twitter.com	35.51%
reportsecret.com	7.02%	youtube.com	4.21%
riafan.ru	3.42%	vine.co	3.94%
politexpert.net	2.10%	factissues.com	3.24%
youtube.com	1.88%	blogspot.com.cy	1.92%
vk.com	1.58%	instagram.com	1.90%
instagram.com	1.53%	facebook.com	1.68%
yandex.ru	1.50%	worldstarr.info	1.47%
infreactor.org	1.36%	trendytopic.info	1.39%
cbslocal.com	1.35%	minibird.jp	1.25%
livejournal	1.35%	yaadlinksradio.com	1.24%
nevnov.ru	1.07%	soundcloud.com	1.24%
ksnt.com	1.01%	linklist.me	1.15%
kron4.com	0.93%	twimg.com	1.09%
viid.me	0.93%	appparse.com	1.08%
newinform.com	0.89%	cargobayy.net	0.88%
inforeactor.ru	0.84%	virralclub.com	0.84%
rt.com	0.81%	tistory.com	0.50%
washigntonpost.com	0.75%	twitcasting.tv	0.49%
seattletimes.com	0.73%	nytimes.com	0.48%

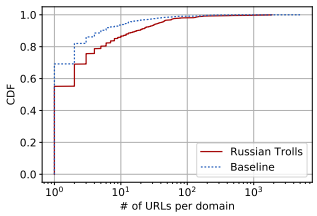
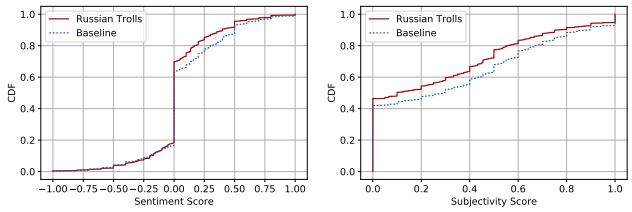


Figure 5: CDF of number of URLs per domain.



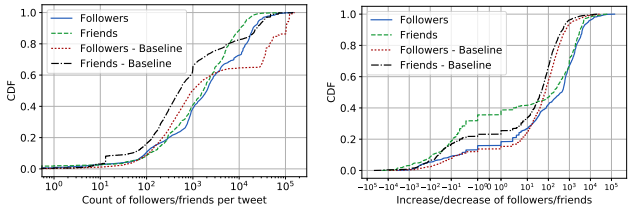
(a) (b)

Figure 6: CDF of sentiment and subjectivity scores for tweets of Russian trolls and random users.

Sentiment analysis. Next, we assess the sentiment and subjectivity of each tweet for both datasets. Fig. 6(a) plots the CDF of the sentiment scores of tweets posted by Russian trolls and our baseline users. We observe that 30% of the tweets from Russian trolls have a positive sentiment, and 18% negative. These scores are not too distant from those of random users where 36% are positive and 16% negative, however, Russian trolls exhibit a unique behavior in terms of sentiment, as a two-sample Kolmogorov-Smirnov test unveils significant differences between the distributions ($p < 0.01$). Overall, we observe that Russian trolls tend to be more negative/neutral, while our baseline is more positive. We also compare subjectivity scores (Fig. 6(b)), finding that tweets from trolls tend to be more subjective; again, we perform significance tests revealing differences between the two distributions ($p < 0.01$).

LDA analysis. We also use the Latent Dirichlet Allocation (LDA) model to analyze tweets' semantics. We train an LDA model for each of the datasets and extract 10 distinct topics with 10 words, as reported in Table 7. Overall, topics from Russian trolls refer to specific world events (e.g., Charlottesville) as well as specific news related to politics (e.g., North Korea and Donald Trump). By contrast, topics extracted from the random sample are more general (e.g., tweets regarding birthdays).

Screen name changes. Previous work [17] has shown that malicious accounts often change their screen name in order to assume different identifies. Therefore, we investigate whether trolls show a similar behavior, as they might change the narrative with which they are attempting to influence public opinion. Indeed, we find that 9% of the accounts operated by trolls change their screen name, up to 4 times during the course of our dataset. Some examples include changing screen names from "OnlineHouston" to "HoustonTopNews," or "Jesus Quintin Perez" to "WorldNewsPolitics," in a clear attempt to pose as news-related accounts. In our baseline, we find that 19% of the accounts changed their Twitter screen names, up



(a) (b)

Figure 7: CDF of the number of (a) followers/friends for each tweet and (b) increase in followers/friends for each user from the first to the last tweet.

to 11 times during our dataset; highlighting that changing screen names is a common behavior of Twitter users in general.

Followers/Friends. Next, we look at the number of followers and friends (i.e., the accounts one follows) of the Russian trolls, as this is an indication of the overall impact of a tweet. In Fig. 7(a), we plot the CDF of the number of followers per tweet measured at the time of that tweet. On average, Russian trolls have 7K followers and 3K friends, while our baseline has 25K followers and 6K friends. We also note that in both samples, tweets reached a large number of Twitter users; at least 1K followers, with peaks up to 145K followers. These results highlight that Russian trolls have a non-negligible number of followers, which can assist in pushing specific narratives to a much greater number of Twitter users. We also assess the evolution of the Russian trolls in terms of the number of their followers and friends. To this end, we get the follower and friend count for each user on their first and last tweet and calculate the difference. Fig. 7(b) plots the CDF of the increase/decrease of the followers and friends for each troll as well as random user in our baseline. We observe that, on average, Russian trolls increase their number of followers and friends by 2,065 and 1,225, respectively, whereas for the baseline we observe an increase of 425 and 133 for followers and friends, respectively. This suggests that Russian trolls work hard to increase their reachability within Twitter.

Tweet Deletion. Arguably, a reasonable strategy to avoid detection after posting tweets that aim to manipulate other users might be to delete them. This is particularly useful when troll accounts change their identity and need to modify the narrative that they use to influence public opinion. With each tweet, the Streaming API returns the total number of available tweets a user has up to that time. Retrieving this count allows us to observe if a user has deleted a tweet, and around what period; we call this an "observed deletion." Recall that our dataset is based on the 1% sample of Twitter, thus, we can only estimate, in a conservative way, how many tweets are deleted; specifically, in between subsequent tweets, a user may have deleted and posted tweets that we do not observe. In Fig. 8, we plot the CDF of the number of deleted tweets per observed deletion. We observe that 13% of the Russian trolls delete some of their tweets, with a median percentage of tweet deletion equal to 9.7%. Whereas, for the baseline set, 27% of the accounts delete at least one tweet, but the median percentage is 0.1%. This means that the trolls delete their tweets in batches, possibly trying to cover their tracks or get a clean slate, while random users make a larger number of deletions but only a small percentage of their overall tweets, possibly

Table 7: Terms extracted from LDA topics of tweets from Russian trolls and baseline users.

Topic	Terms (Trolls)	Topic	Terms (Baseline)
1	trump, black, people, really, one, enlist, truth, work, can, get	1	want, can, just, follow, now, get, see, don, love, will
2	trump, year, old, just, run, obama, breaking, will, news, police	2	2016, july, come, https, trump, social, just, media, jabberduck, get
3	new, trump, just, breaking, obamacare, one, sessions, senate, politics, york	3	happy, best, make, birthday, video, days, come, back, still, little
4	man, police, news, killed, shot, shooting, woman, dead, breaking, death	4	know, never, get, love, just, night, one, give, time, can
5	trump, media, tcot, just, pjnet, war, like, video, post, hillary	5	just, can, everyone, think, get, white, fifth, veranomtv2016, harmony, friends
6	sports, video, game, music, isis, charlottesville, will, new, health, amb	6	good, like, people, lol, don, just, look, today, said, keep
7	can, don, people, want, know, see, black, get, just, like	7	summer, seconds, team, people, miss, don, will, photo, veranomtv2016, new
8	trump, clinton, politics, hillary, video, white, donald, president, house, calls	8	like, twitter, https, first, can, get, music, better, wait, really
9	news, world, money, business, new, one, says, state, 2016, peace	9	dallas, right, fuck, vote, police, via, just, killed, teenchoice, aldubmainecelebration
10	now, trump, north, korea, people, right, will, check, just, playing	10	day, black, love, thank, great, new, now, matter, can, much

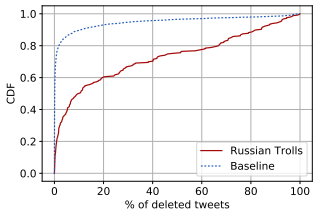


Figure 8: CDF of the number of deleted tweets per observe deletion.

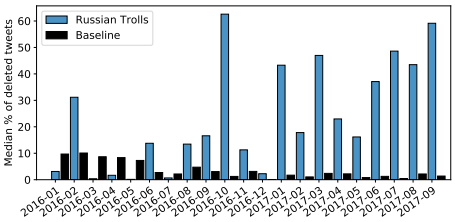


Figure 9: Average percentage of observed deletions per month.

because of typos. We also report the distribution, over each month, of tweet deletions in Fig. 9. Specifically, we report the mean of the percentages for all observed deletions in our datasets. Most of the tweets from Russian trolls are deleted in October 2016, suggesting that these accounts attempted to get a clean slate a few months before the 2016 US elections.

Case Study. While the previous results provide a quantitative characterization of Russian trolls behavior, we believe there is value showing a concrete example of the behavior exhibited and how techniques played out. We start on May 15, 2016, where the troll with screen name 'Pen_Air', was posing as a news account via its profile description: "National American news." On September 8, 2016 as the US presidential elections approached, 'Pen_Air' became a Trump supporter, changing its screen name to 'Blacks4DTrump' with a profile description of "African-Americans stand with Trump to make America Great Again!" Over the next 11 months, the account's tweet count grew from 49 to 642 while its follower count rose from 1.2K to *nearly* 9K. Then, around August 18, 2017, the account was seemingly repurposed. Almost all of its previous tweets were deleted (the account's tweet count dropped to 35), it gained a new screen name ('southlonestar2'), and was now posing as a "Proud American and TEXAN patriot! Stop ISLAM and PC. Don't mess with Texas" according to its profile description. When examining the accounts tweets, we see that most are clearly related

Table 8: Total URLs with at least one event in Twitter, /pol/, Reddit, and Russian trolls on Twitter; total events for Russian state-sponsored news URLs, other news URLs and all the URLs; and mean background rate (λ_0) for each platform.

		/pol/	Reddit	Twitter	Trolls
URLs	Russian state-sponsored	6	13	19	19
	Other news sources	47	168	159	192
	All	127	482	861	989
Events	Russian state-sponsored	19	42	118	19
	Other news sources	720	3,055	2,930	195
	All	1,685	9,531	1,537,612	1,461
Mean λ_0	Russian state-sponsored	0.0824	0.1865	0.2264	0.1228
	Other news sources	0.0421	0.1447	0.1544	0.0663
	All	0.0324	0.1557	0.1553	0.0753

to politics, featuring blunt right-wing attacks and "talking points." For example, "Mr. Obama! Maybe you bring your girls and leave them in the bathroom with a grown man! #bathroombill #NObama <url>" on May 15, 2016, "#HiLIARy has only two faces! And I hate both! #NeverHillary #Hillaryliesmatter <url>" on May 19, 2016, and "RT @TEN_GOP: WikiLeaks #DNCLeaks confirms something we all know: system is totally rigged! #NeverHillary <url>." on July 22, 2016.

Take-aways. In summary, our analysis leads to the following observations. First, we find evidence that trolls were actively involved in the dissemination of content related to world news and politics, as well as propaganda content regarding various topics such as ISIS and Islam. Moreover, several Russian trolls were created or repurposed a few weeks before notable world events, including the Republican National Convention meeting or the Charlottesville rally. We also find that the trolls deleted a substantial amount of tweets in batches and overall made substantial changes to their accounts during the course of their lifespan. Specifically, they changed their screen names aiming to pose as news outlets, experienced significant rises in the numbers of followers and friends, etc. Furthermore, our location analysis shows that Russian trolls might have tried to manipulate users located in the USA, Germany, and possibly in their own country (i.e., Russia), by appearing to be located in those countries. Finally, the fact that these accounts were active up until their recent suspension also highlights the need to develop more effective tools to detect such actors.

4 INFLUENCE ESTIMATION

Thus far, we have analyzed the behavior of the Russian trolls on the Twitter platform, and how this differs from that of a baseline of random users. Allegedly, their main goal is to ultimately manipulate the opinion of other users and extend the cascade of disinformation

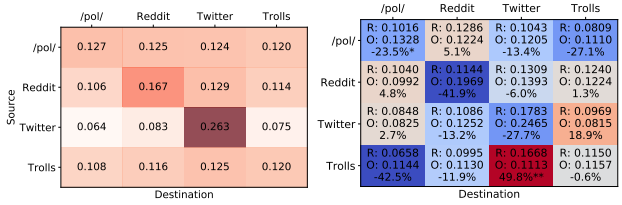
they share (e.g., other users post similar content) [4]. Therefore, we now set out to shed light on their impact, in terms of the dissemination of disinformation, on Twitter and on the greater Web.

To assess their influence, we look at the URLs posted by four groups of users: Russian trolls on Twitter, "normal" accounts on Twitter, Reddit users, and 4chan users (/pol/ board). For each unique URL, we fit a statistical model known as Hawkes Processes [15, 16], which allows us to estimate the strength of connections between each of these four groups in terms of how likely an event – the URL being posted by either trolls or normal users to a particular platform – is to cause subsequent events in each of the groups. For example, a strong connection from Reddit to /pol/ would mean that a URL that appears on Reddit is likely to be seen and then re-posted on /pol/; whereas, a weak connection from trolls to normal users on Twitter indicates that a URL posted by trolls is less likely to be re-tweeted or re-posted by the latter. We fit the Hawkes Processes using the methodology presented by [29].

To study the dissemination of different types of content, we look at three different sets of URLs: 1) The complete set of all URLs posted by Russian trolls; 2) The subset of URLs for Russian state-sponsored news, namely, RT (Russia Today); and 3) The subset of URLs from other mainstream and alternative news sources using the list provided by [29]. Table 8 summarizes the number of URLs, number of events (i.e., occurrences of a given URL) as well as the mean background rate for each category and social network. The background rate defines the rate at which events occur excluding the influence of the platforms included in the model; the background rate includes events created spontaneously on each platform, such as by a user sharing the article from the original source, or those generated by another platform not monitored by us like Facebook. The number of events for Russian state-sponsored news sources is substantially lower than the number of events from other news sources. This is expected since the former only includes one news source (RT), however, it is interesting that the background rates for these URLs are higher than for other news sources, meaning that events from Russian state-sponsored news are more likely to occur spontaneously.

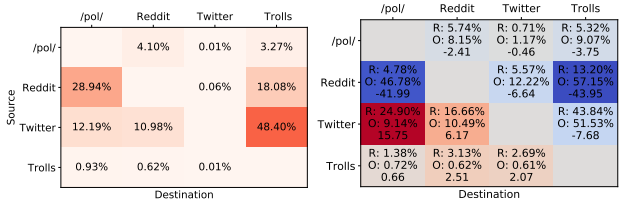
Fitting a Hawkes model yields a weight matrix, which characterizes the strength of the connections between the groups we study. Each weight value represents the connection strength from one group to another and can be interpreted as the expected number of subsequent events that will occur on the second group after each event on the first. The mean weight values over all URLs, as well as for the URLs from RT and other news URLs, are presented in Fig. 10. We observe that for /pol/, Reddit, and normal users on Twitter, the greatest weights are from each group to itself, meaning that reposts/retweets on the same site are more common than sharing the URL to the other platforms (Fig. 10(a)). For the Russian Trolls on Twitter, however, the weight is greater from the trolls to Twitter than from the trolls to themselves, perhaps reflecting their use as an avenue for *disseminating* information to normal Twitter users (Fig. 10(b)). Also, we observe that, in most cases, the connections are stronger for non-Russia state-sponsored news, indicating that regular users are more inclined to share news articles from mainstream and alternative news sources.

Looking at the Russian trolls and normal Twitter users, we see that the trolls are more likely to retweet Russian state-sponsored



(a) All URLs (b) News URLs

Figure 10: Mean weights for (a) all URLs in our dataset and (b) news URLs categorized as Russian state-sponsored (R) and other mainstream and alternative news URLs (O). We also show the percent of increase/decrease between the two categories. Note that * and ** refer to statistical significance with, resp., $p < 0.05$ and $p < 0.01$.



(a) All URLs (b) News URLs

Figure 11: Estimated mean percentages of events created because of other events for (a) all URLs and (b) Russian state-sponsored URLs (R) and other mainstream and alternative news URLs (O). We also show the difference between the two categories of news.

URLs from normal Twitter users than other news sources; conversely, normal Twitter users are more likely to retweet Russian state-sponsored URLs from the troll accounts. In order to assess the significance of our results, we perform two-sample Kolmogorov-Smirnov tests on the weight distributions for the RT URLs and the other news URLs for each source-destination platform pair (depicted as stars in the Fig. 10(b)). Small p value means there is a statistically significant difference in the way that RT URLs propagate from the source to the destination platform. Most of the source-destination pairs have no statistical significance, however for the Russian trolls–Twitter users pair, we find significance difference with $p < 0.01$.

In Fig. 11, we report the estimated total impact for each pair of platforms, for both Russian state-sponsored news, other news sources as well as all the observed URLs. We determine the impact by calculating, based on the estimated weights and the number of events, the percentage of events on a destination platform caused by events on a source platform, following the same methodology as [29]. For all URLs (Fig. 11(a)), we find that the influence of Russian trolls is negligible on Twitter (0.01%), while for /pol/ and Reddit it is slightly higher (0.93% and 0.62%, respectively). For other pairs, the larger impacts are between Reddit–/pol/ and Twitter-Russian trolls, mainly due to the larger population of users. Looking at the estimated impact for RT and other news sources (Fig. 11(b)), we note that the trolls influenced the other platforms approximately the same for alternative and mainstream news sources (0.72%, 0.62%, and 0.61 for /pol/, Reddit, and Twitter, respectively). Interestingly,

Russian trolls have a much larger impact on all the other platforms for the RT news when compared to the other news sources: approximately 2 times more on /pol/, 5 times more on Reddit, and 4 times more on Twitter.

Take-aways. Using Hawkes processes, we were able to assess the degree of influence Russian trolls had on Twitter, Reddit, and /pol/ by examining the diffusion of information via URLs to news. Our results indicate that their influence is actually quite limited. With the exception of news originating from the Russian state-sponsored news outlet RT, the troll accounts were generally less influential than other users on Reddit, Twitter, and 4chan. However, our analysis is based only on 1K troll accounts found in Twitter's 1% stream, and, as mentioned previously, Twitter recently announced they had discovered over 1K more trolls and more than 50K automated accounts. With that in mind, there are several potential explanations behind this limited influence. For example, there might be a lot of influence attributed to regular Twitter users that belongs to newly announced troll accounts. Considering that Twitter announced the discovery of "only" 1K more troll accounts, we suspect that this is not really the case. Another, more plausible explanation is that the troll accounts are just not terribly efficient at spreading news, and instead are more concerned with causing havoc by pushing ideas, engaging other users, or even taking both sides of controversial online discussions [22]. This scenario makes more sense considering that, along with 1K new troll accounts, Twitter also announced discovering over 50K *automated* accounts that might be more efficient in terms of driving traffic to specific URLs.

5 RELATED WORK

Opinion manipulation. The practice of swaying opinion on the Web is a long-standing issue as malicious actors attempt to push their agenda. Kumar et al. [13] study how users create multiple accounts, called *sockpuppets*, that participate in Web communities to manipulate users' opinions. They find that sockpuppets exhibit different posting behavior when compared to benign users. Mihaylov et al. [18] show that trolls can manipulate users' opinions in forums, while in their follow-up work [19] they highlight the two types of manipulation trolls: those paid to operate and those that are called out as such by other users. Then, Volkova and Bell [26] predict the deletion of Twitter accounts because they are trolls, focusing on those that shared content related to the Russian-Ukraine crisis.Elyashar et al. [6] distinguish authentic discussions from campaigns to manipulate the public's opinion, using a set of similarity functions alongside historical data. Also, Steward et al. [22] focus on discussions related to the Black Lives Matter movement and how content from Russian trolls was retweeted by other users. Using the retweet network, they find the existence of two communities; one left- and one right-leaning communities. Also, they note that trolls infiltrated both communities, setting out to push specific narratives. Finally, Varol et al. [25] aim to identify memes that become popular due to *coordinated* efforts, and achieve 75% AUC score before memes become trending and 95% AUC score afterwards.

False information on the political stage. Conover et al. [2] study the interactions of right- and left-leaning communities on Twitter during the 2010 US midterm elections: finding that the retweet network has limited connectivity between the two communities, which does not happen in the mentions network; mainly because users engage others users with different ideologies and expose them to different opinions. Ratkiewicz et al. [20] use machine learning to detect the early stages of false political information spreading on Twitter and introduce a framework that considers topological, content-based, and crowdsourced features of the information diffusion. Wong et al. [27] quantify political leaning of users and news outlets during the 2012 US presidential election on Twitter by using an inference engine that considers tweeting and retweeting behavior of articles. Yang et al. [28] investigate the topics of discussions on Twitter for 51 US political persons showing that Democrats and Republicans are active in a similar way on Twitter. Le et al. [14] study 50M tweets regarding the 2016 US election primaries and highlight the importance of three factors in political discussions on social media, namely the *party*, *policy considerations*, and *personality* of the candidates. Howard and Kollanyi [12] study the role of bots in Twitter conversations during the 2016 Brexit referendum. By analyzing 1.5M tweets, they find that most tweets are in favor of Brexit and that there are bots with various levels of automation.Also, Hegelich and Janetzko [10] highlight that bots have a political agenda and that they exhibit various behaviors, e.g., trying to hide their identity and promoting topics through hashtags and retweets. Finally, a large body of work focuses on social bots [1, 3, 7, 8, 24] and their role in spreading disinformation, highlighting that they can manipulate the public's opinion at large scale, thus potentially affecting the outcome of political elections.

Remarks. Unlike previous work, our study focuses on the set of Russian troll accounts that were suspended by Twitter and released by the US congress. To the best of our knowledge, this constitutes the first effort not only to characterize a ground truth of troll accounts independently identified by Twitter, but also to quantify their influence on the greater Web, specifically, on Twitter as well as on Reddit and 4chan.

6 CONCLUSION

In this paper, we analyzed the behavior and use of the Twitter platform by Russian trolls during the course of 21 months. We showed that Russian trolls exhibited interesting differences when compared with a set of random users, actively disseminated politics-related content, adopted multiple identities during their account's lifespan, and that they aimed to increase their impact on Twitter by increasing their followers. Also, we quantified the influence that Russian trolls have on Twitter, Reddit, and /pol/ using a statistical model known as Hawkes Processes. Our findings show that trolls' influence was not substantial with respect to the other platforms, with the significant exception of news published by the Russian state-sponsored news outlet RT.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska Curie ENCASE project (GA No. 691025) and under the CYBERSECURITY CONCORDIA project (GA No. 830927).

REFERENCES

[1] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).

[2] Michael Conover, Jacob Ratkiewicz, Matthew R. Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

[3] Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. In *WWW*.

[4] Samuel Earle. 2017. Trolls, Bots and Fake News: The Mysterious World of Social Media Manipulation. https://goo.gl/nz7E8r.

[5] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2017. Towards detecting compromised accounts on social networks. *IEEE TDSC* (2017).

[6] Aviad Elyashar, Jorge Bendahan, and Rami Puzis. 2017. Is the Online Discussion Manipulated? Quantifying the Online Discussion Authenticity within Online Social Media. *CoRR* abs/1708.02763 (2017).

[7] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *ArXiv 1707.00086* (2017).

[8] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* (2016).

[9] Aditi Gupta and Ponnurangam Kumaraguru. 2012. Credibility ranking of tweets during high impact events. In *PSOSM '12*.

[10] Simon Hegelich and Dietmar Janetzko. 2016. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In *ICWSM*.

[11] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.

[12] Philip N. Howard and Bence Kollanyi. 2016. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *Arxiv 1606.06356* (2016).

[13] Srijan Kumar, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *WWW*.

[14] Huyen T. Le, G. R. Boynton, Yelena Mejova, Zubair Shafiq, and Padmini Srinivasan. 2017. Revisiting The American Voter on Twitter. In *CHI*.

[15] Scott W. Linderman and Ryan P. Adams. 2014. Discovering Latent Network Structure in Point Process Data. In *ICML*.

[16] S. W. Linderman and R. P. Adams. 2015. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv 1507.03228* (2015).

[17] Enrico Mariconti, Jeremiah Onaolapo, Syed Sharique Ahmad, Nicolas Nikiforou, Manuel Egele, Nick Nikiforakis, and Gianluca Stringhini. 2017. What's in a Name?: Understanding Profile Name Reuse on Twitter. In *WWW*.

[18] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding Opinion Manipulation Trolls in News Community Forums. In *CoNLL*.

[19] Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. In *ACL*.

[20] Jacob Ratkiewicz, Michael Conover, Mark R. Meiss, Bruno GonÃğalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and Tracking Political Abuse in Social Media. In *ICWSM*.

[21] Kate Starbird. 2017. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *ICWSM*.

[22] Leo Steward, Ahmer Arif, and Kate Starbird. 2018. Examining Trolls and Polarization with a Retweet Network. In *MIS2*.

[23] The Independent. 2017. St Petersburg 'troll farm' had 90 dedicated staff working to influence US election campaign. https://ind.pn/2yuCQdy.

[24] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *ICWSM*.

[25] Onur Varol, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2017. Early detection of promoted campaigns on social media. *EPJ Data Science* (2017).

[26] Svitlana Volkova and Eric Bell. 2016. Account Deletion Prediction on RuNet: A Case Study of Suspicious Twitter Accounts Active During the Russian-Ukrainian Crisis. In *NAACL-HLT*.

[27] Felix Ming Fai Wong, Chee-Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*.

[28] Xinxin Yang, Bo-Chiuan Chen, Mrinmoy Maity, and Emilio Ferrara. 2016. Social Politics: Agenda Setting and Political Communication on Social Media. In *SocInfo*.

[29] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM IMC*.

Data Science for Social Good Chairs' Welcome

It is our great pleasure to welcome you to the *WWW 2019 Data Science for Social Good Workshop.*

The need for social innovation is clear: From climate change to income inequality to geopolitical upheaval and terrorism, the difficulties confronting us are unprecedented not only in their variety but also in their complexity. At the same time, today's public policy practices and tools are not sufficient. Increasingly, it is clear, we need not only new solutions but new methods to arriving at solutions. Data and data science will become more central to meeting these challenges and to social innovation, philanthropy, international development and humanitarian aid. From the analysis of satellite imagery to mapping poverty to using Facebook data to track the global digital gender gap, "Data Science for Social Good" provides great promise. Data from corporate actors (e.g. mobile phones data, remote sensing, satellite imagery) as well as data from digital traces generated by the pervasiveness of the Web in combination with state-of-the-art knowledge generated by data science can be synergically exploited to solve issues around many social problems and support global agencies and policymakers in implementing better and more impactful policies and interventions.

Yet, for all of data's potential to address public challenges, the truth remains that most of the data assets and data science capabilities available today are not yet sufficiently applied to solving public problems. Because of a lack of awareness of the potential, funding and data access restrictions and often poorly distributed data science capacity, its vast potential often goes untapped.

This workshop will review the potential and emerging field of data science for good – as well as how to develop new partnerships for the data age (Data Collaboratives) to unlock both data and data science capabilities.

This Workshop is an example of Data Collaboratives, a new form of collaboration, beyond the public-private partnership model, in which participants from different sectors — especially companies — exchange their data to create public value. Also, the Workshop particularly welcomed those contributions which aim at closing the gap between developing models and algorithms, and their practical applications in the workflow of non-profit organizations.

Key questions we will seek to address through this workshop include:

- What are the value propositions and practices of data science for social good?
- What are different ways to engage data scientists (e.g. through data collaboratives) - examples and lessons learned?
- What are the risks of using data science for social good?
- What data science methodologies are already available or need to be accelerated to solve public problems?
- How to overcome the current transactions costs and funding constraints associated with leveraging data and data science for social good?

Subject areas:

- International development
- Humanitarian aid
- Gender data gaps
- Health in developing countries
- Migrations
- Education

- Unemployment
- Inequality and poverty reduction
- Environment and sustainability

We have received several very interesting submissions from different countries. We have decided to accept 6 papers and 4 abstracts. These works cover a broad range of topics: Human Values and Attitudes towards Vaccination; Crosslingual Perspectives about Crisis-related Conversations on Twitter; Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US midterms; Spatial-Temporal Correlation Modeling of Community-Based Air Monitorin Sensor Network; Detecting Areas of Potential High Prevalence of Chagas in Argentina; Creating Public Value by Democratizing the Ecosystem of Human Service Providers; Crowdsourced Community Health Surveillance during Emergencies using Web-based Cloud Technology; Policy Priority Inference: a Computational Approach for the Discovery, Evaluation and Prescription of Development Strategies; Fairsearch: A programming Library for Fair Search Results.

We also succeeded in attracting four prominent speakers and three outstanding panelists to discuss about data science for social good using private data.

We believe that the Workshop can be an important occasion where researchers from the Web Conference Community as well as from the fields of data science, machine learning and artificial intelligence together with experts in the social and political sciences will be able to present and discuss applications of data science with a high social impact; and will examine why, where, how and under what conditions data science can advance social objectives.

Data Science for Social Good Organization

Workshop/Track Chairs:
Natalia Adler, UNICEF, NY, USA
Ciro Cattuto, ISI Foundation, Turin, Italy
Daniela Paolotti, ISI Foundation, Turin, Italy
Michele Tizzoni, ISI Foundation, Turin, Italy
Stefaan Verhulst, GovLab, NY, USA
Andrew Young, GovLab, NY, USA

Program Committee:
Luca Maria Aiello, Nokia Bell Labs, London, UK
Manuel Garcia-Herranz, UNICEF Innovation, New York, USA
Eduardo Graells-Garrido, Universidad del Desarrollo, Santiago de Chile, Chile
Kyriaki Kalimeri, ISI Foundation, Turin, Italy
Miguel Luengo-Oroz, UN Global Pulse, New York, USA
Yelena Mejova, ISI Foundation, Turin, Italy
Elisa Omodei, UNICEF Innovation, New York, USA
Jose Ramasco, University of Balearic Islands, Palma de Mallorca, Spain
Rossano Schifanella, University of Turin, Turin, Italy

Problems and Opportunities of Working with a Telco's Large Data Sets of Mobile Data*

Leo Ferres
Data Science Institute
Universidad del Desarrollo, Santiago, Chile; Telefónica Research andDevelopment, Santiago, Chile, lferres@udd.cl

ABSTRACT

Since 2016, through an association between Telefónica R&D and the Institute of Data Science in Chile, a group of researchers and myself have been working with trillions of digital traces left behind when people use their mobile phones. All of this work has been done under the general umbrella term of "data science for social good", and we have worked on anything from population displacement after external events like earthquakes, how people started using public spaces after the introduction of a popular mobile game, to actual social inclusion of people of different socio-economic backgrounds mixing in shopping malls or reading certain kinds of news, or patterns arising from gendered data sets. We will show how data in the private sector made us learn important social lessons such as how parks can become more secure when people went out to play Pokemon Go, how certain malls are hubs of social inclusion, how gender segregates the city and how different demographics keep themselves in their own informational filter bubble. However, even after all this benefits, the relationship with industry has never been fluid, and involves a lot of small and not so small compromises and "battles". In this talk, I will present a technical history of the work we've done with X/CDRs for social good including practical aspects of accessing and sharing data, the balance of research and industrial innovation, and issues of transactions costs while still providing value for the company itself, government, the university and society. I will also recount experiences about what it meant for a company like Telefónica and a research university like us to travel together in a very interesting context of huge data, incredible insights, privacy considerations, money, corporate interests, university expectations, and data-driven discovery.

CCS CONCEPTS

• General and Reference

KEYWORDS

Mobility, CDR, gender

ACM Reference format:

Leo Ferres. 2019. Problems and Opportunities of Working with a Telco's Large Data Sets of Mobile Data. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316490

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316490

AI for Good at Scale: Challenges in Real Time Emergency Response

Alex Jaimes
AI & Data Science division
DataMinr, New York, USA, aj27@caa.columbia.edu

ABSTRACT

In most emergencies, people use social media platforms to publicly share information. Such data, from multiple sources is extremely useful for emergency response and public safety: the more knowledge that is gathered, the better the response can be. When an emergency event first occurs, getting the right information as quickly as possible is critical in saving lives. When an emergency event is ongoing, information on what is happening can be critical in making decisions to keep people safe and take control of the particular situation unfolding. In both cases, first responders have to quickly make decisions that include what resources to deploy and where. In this talk, I will describe challenges in emergency response and how a computational platform that leverages public data can address them. A platform to detect emergency situations and deliver the right information to first responders has to deal with ingesting thousands of data points per second: sifting through and identifying relevant information, from different sources, in different formats, with varying levels of detail, in real time, so that first responders and others can be alerted at the right level and at the right time. I will describe technical challenges in processing vast amounts of heterogenous data in real time, highlighting the importance of interdisciplinary research and a human-centered approach to address problems in emergency response. I will give specific examples and discuss relevant research topics in Machine Learning, NLP, Information Retrieval, Computer Vision and other fields.

CCS CONCEPTS

• General and Reference

KEYWORDS

Emergency, data, NLP

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316492

ACM Reference format:

Alex Jaimes. 2019. AI for Good at Scale: Challenges in Real Time Emergency Response. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316492

Creating Public Value by Democratizing the Ecosystem of Human Service Providers

Charalampos Chelmis
University at Albany, SUNY
Department of Computer Science
cchelmis@albany.edu

Mengfan Yao
University at Albany, SUNY
Department of Computer Science
myao@albany.edu

ABSTRACT

Human service providers play a critical role in improving well–being in the United States. However, little is know about (i) how service seekers find the services they are looking for by navigating among available service providers, and (ii) how such organizations collaborate to meet human needs. In this paper, we report the first outcomes of our ongoing project. Specifically, we first describe a data acquisition engine, designed around the particular challenges of capturing, maintaining, and updating data pertaining to human service organizations from semistructured Web sources. We then proceed to illustrate the potential of the resulting comprehensive repository of human service providers through a case study showcasing a mobile app prototype designed to provide a one–stop shop for human service seekers.

CCS CONCEPTS

• **Information systems** → **Web crawling**; *Information systems applications*; • **Human-centered computing** → *Graphical user interfaces.*

KEYWORDS

Web and Society; Web crawling; Human services; Non–profit organizations; Applied data science

ACM Reference Format:
Charalampos Chelmis and Mengfan Yao. 2019. Creating Public Value by Democratizing the Ecosystem of Human Service Providers. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316487

1 INTRODUCTION

Navigating among service providers is an essential part of seeking for help [2, 7, 8, 11, 16, 20, 23, 25]. And yet, the process of seeking and providing services is often slow and fragmented: (i) service providers often work in silos [4, 10, 14, 15], whereas (ii) people who seek help carry the burden of navigating through available services and eligibility criteria [2]. In our ongoing project, we focus on the world of human service providers [12] with the ultimate goal of understanding how service seekers find the services they are looking for. Unlike the Web where people can find and access relevant information by reviewing websites or navigating webpages by following hyperlinks, no comprehensive resource exists to date for service seekers to find service providers. Additionally, the network structure (if any) among service providers is unknown. Thus, service seekers must rely on the information they already have (e.g., list of food pantries in an area) or new information they can obtain (i.e., through referrals) in order to eventually reach a service provider that can cater to their needs.

To understand how human service providers come together to address the grand challenges associated with meeting human needs, the most crucial problem to solve is finding an appropriate data source. To overcome this barrier, we leverage the medium that most (if not all) human service organizations rely upon for their public communication, i.e., their websites. The advantage of using websites to collect information about human service organizations is twofold. First, websites are publicly available online and their content is machine–readable. Second, they provide rich and semi–structured data that can be used to develop a comprehensive database of organizations and their partners, which can in turn be subsequently used to (i) develop technologies that enable novel capabilities, such as the one described in this paper, and (ii) analyze the interactions among human service organizations.

The limitation of using Web data is that it may lead to an imperfect representation of the actual set of human service organizations and their connections. For instance, organizations may choose to (i) not have a web presence at all, (ii) not list their partners on their websites, or (iii) use offline resources such as printed books similar to yellow pages to connect to other organizations. From the perspective of a service seeker however, information available on the Web may be the only resource to find services due to reasons, including but not limited to, restricted mobility (e.g., a homeless person may be unable to travel the distance to a referral organization so as to ask questions) or personal preferences (e.g., someone may prefer the anonymity of the Web to investigate for available options). Even though a subsequent study to understand the quality of this approximation may be necessary, it is beyond the scope of this work.

We begin by providing the key goal of our ongoing project (Section 2), and subsequently describe our data acquisition engine (and associated challenges) to collect the greatest amount of potentially useful information possible (including the mission and scope of an organization, hours of operation and list of services provided) about human service providers in a given locale (Section 4). We then illustrate the potential of the resulting repository of human service providers to create **public value** through a mobile app for

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316487

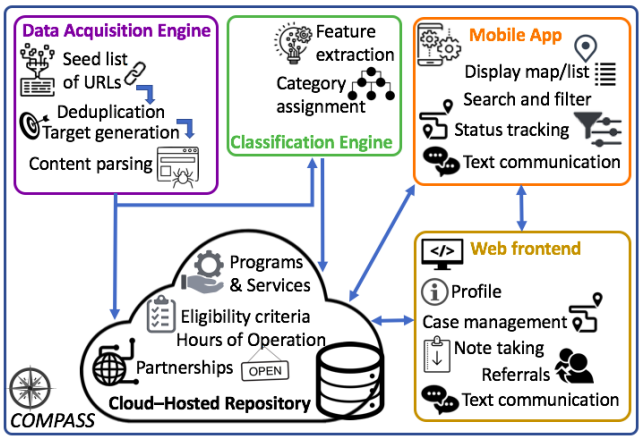


Figure 1: System architecture.

searching and filtering service providers, and requesting and tracking services through the click of a button (Section 5). We conclude with a discussion and future work in Section 6.

2 BACKGROUND

The principal objective of our ongoing project is threefold: (i) democratize access to the greatest amount of potentially useful information about available human services in a locale, without overwhelming the user with excess information, (ii) seamlessly connect service seekers to providers and facilitate digital interactions, and (iii) provide service seekers with a way to track the whole service cycle, from the moment a request is initiated, to service delivery, to outcome(s), all through a user–friendly, easy to use mobile app.

To accomplish this ambitious objective, we are currently developing *COMPASS*, a system comprising five modules: (i) data acquisition engine, (ii) classification engine, (iii) comprehensive repository and information integration engine, (iv) Web frontend, and (v) mobile app, as illustrated in Figure 1. In this paper, we present our progress on the data acquisition engine (Section 4) and the mobile app (Section 5) components of our framework.

3 RELATED WORK

Data collection from the Web. A large number of general–purpose crawlers have thus far been developed for scraping information from online sources [1, 13, 21]. While such crawlers are designed to collect data at scale, retrieving information about human services by scraping publicly available information from their websites presents unique challenges (see Section 4).

Technologies for Underserved Populations and Non–profit Organizations. Nearly 70% of low–income consumers rely primarily on their smartphones for connecting to the Web [9, 24]. Therefore, numerous mobile apps (e.g., Range[1] and OneToday[2]) have been developed to assist underserved populations get access to information, such as finding the nearest grocery store that accepts food stamps or the best prices for milk in a given area [3]. Apps specifically designed to address food insecurity and homelessness include [5, 6, 17, 18]. The use of technology in non–profit organizations has been explored in [27].

4 DATA ACQUISITION ENGINE

The goal of our data acquisition engine is to collect the greatest information possible about as many human service providers as possible within a given locale. In our current work, we focus on the metropolitan area surrounding Albany, the capital of the U.S. state of New York. Albany, the 4th largest metropolitan region in the state and the 45th largest in the U.S., is a multiracial and multiethnic city that contains nearly 99,000 residents (on average 52% White, 29% Black or African American, 9% Hispanic or Latino, 6% Asian, and growing refugee populations)[3] and a high concentration of human service providers and service organizations related to government, health care, and education [19]. The concentration of human service providers in Albany extends to two adjacent cities, Troy and Schenectady, which exist within a 20–mile radius from Albany. Albany and these two cities form the core of the Capital District of the New York State (officially defined as the Albany–Troy–Schenectady Metropolitan Statistical Area with a population size of approximately 800,000). Although these cities function as independent localities, they share commuting populations, local resources, and services. Nevertheless, our data acquisition engine is not restrained to the specific geographic boundaries of the Capital District, but is instead applicable to any location across the US.

We implemented our data acquisition engine in Python using BeautifulSoup[4] to parse HTML pages. A NoSQL cloud database[5] is used to store the data. The benefit of using a cloud–hosted database is security, persistency, flexibility, and real–time synchronization with every connected mobile app client.

4.1 Completeness

Ideally, a number of attributes, including the mission, physical address, hours of operation, programs and services provided, eligibility criteria, and category (as recorded by the National Taxonomy of Exempt Entities system,[6] which facilitates the categorization of organizations by their types and activities, promotes uniformity and comparability, and is used by the IRS to classify non–profit organizations), in addition to the website address and name of each organization, should be collected and stored in a comprehensive repository. Note that within the major group of "Human Services", the National Taxonomy of Exempt Entities system further divides organizations into logical subdivisions, the first few of which are shown in Figure 2.

Our data acquisition engine visits all webpages originating the starting domain of a given URL, parsing their content for the desired information. Note that we currently rely on the URL of an organization to identify it as unique; however, different URLs may resolve to the same webpage (e.g., due to parameters appended to the URL which do not change the content of the page). On the other hand, some organizations may use the same domain to host their

[1] http://www.rangeapp.org/
[2] https://onetoday.google.com/
[3] http://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2014/5-year.html
[4] https://www.crummy.com/software/BeautifulSoup/
[5] https://firebase.google.com/
[6] https://nccs.urban.org/classification/national-taxonomy-exempt-entities

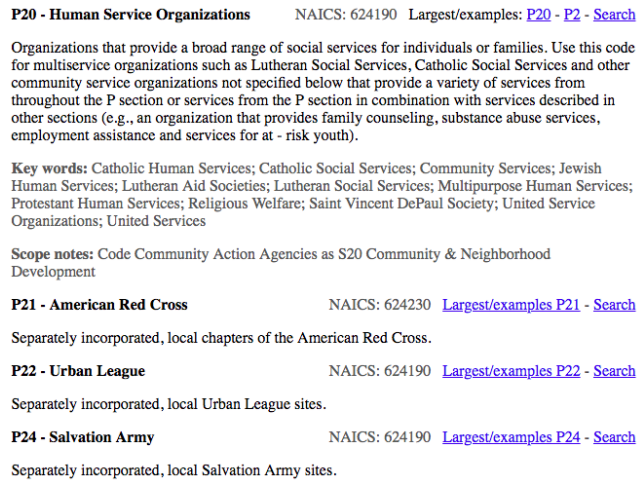


Figure 2: Sample subdivisions of human services according to the National Taxonomy of Exempt Entities system.

webpages. To address this challenge, we perform post–analysis to identify URLs resolving to the same organization. Finally, our data acquisition engine searches for partner organizations while ignoring links to social media platforms such as Facebook and Twitter so that no websites outside the realm of human service organizations are considered. If a domain has a subdomain, the subdomain is also examined.

Additionally, we would like our crawler to collect information about as many organizations in a region as possible. Our data acquisition engine performs snowball sampling, i.e., examines organizations that are listed as partners on websites of organizations in our seed list, to further expand the original list. The outcome depends to a large extent to the initial set of organizations provided as input to our data acquisition engine. Currently, our seed list comprises a total of over $3,000$ service–providing organizations and their corresponding URLs. This list was automatically compiled using Selenium[7] from (i) GuideStar,[8] an information service specializing in reporting on U.S. non–profit organizations, (ii) GreatNonprofits,[9] a website that allows donors, volunteers, and clients to share their personal experiences and reviews of charitable organizations, essentially providing crowdsourced information about the reputability of these organizations, and (iii) Charity Navigator,[10] an independent charity watchdog that evaluates charitable organizations in the United States.

4.2 Flexibility

Each organization's website has its own structure, which in addition may also change over time. Some organizations may have minimalistic websites containing a general overview of their mission and activities, while others might have multiple pages for specific programs and services. With respect to identifying webpages listing partnering organizations, these might appear in various formats, such as in menu items, images, PDF files, or unstructured text. The data acquisition engine should be able to detect and annotate the type of content describing partnering organizations, and either parse it automatically or log it and inform a system administrator to allow for post–processing and/or adaptation of the pertaining module. An easy solution would be to automatically filter navigation menus to identify pages that contain information about partner organizations. However, given differences and lack of standards in web design, this may be infeasible when crawling a number of heterogeneous websites.

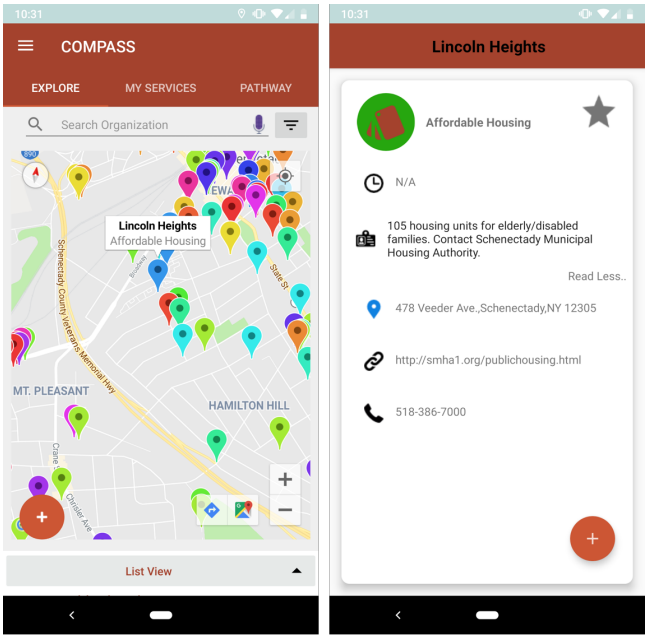
To address such difficulties, our data acquisition engine follows a modular design where its main functionality is implemented in generic modules (e.g., for the interaction with the cloud–hosted database), which do not require modification when adding new URLs for crawling. Our main content parsing module traverses a given domain searching for web pages containing information regarding the mission statement of organizations. Our partners detection module traverses a given domain searching for web pages containing at least three URLs with a ".org" suffix, as indicative of human service organizations, and therefore a potentially good heuristic to identify webpages that list partners. In our ongoing work, we are investigating the use of additional (and/or more appropriate) heuristics in order to enhance the ability of our engine to parse additional formats so as to better capture existing relationships between organizations. When a website uses a previously unseen structure, a new scraping module (i.e., set of regular expression matching operations to identify specific patterns) may be required.

4.3 Efficiency

Information about each parsed organization should be stored in our comprehensive repository exactly once, and the same organization should not be visited more than once during a crawl. In our current implementation, we perform a basic edit distance string similarity check on organizations' title in addition to examining their corresponding URLs to identify and avoid parsing and storing duplicate entries. Our engine additionally provides the means to revisit a website, inspect it, and if needed, update the entry in our repository with new or modified information about the programs and services and partnering organizations of the corresponding service provider. A log system helps to keep track of the crawling status and detect unexpected errors (e.g., site is down) or encountered problems (e.g., network connectivity issues). Most errors and warnings can be handled automatically (e.g., connectivity problems), however some issues require manual inspection (e.g., when a website uses a previously unseen structure). Self–recovery capabilities are used to detect when an error has occurred, and if possible, continue with the data collection process.

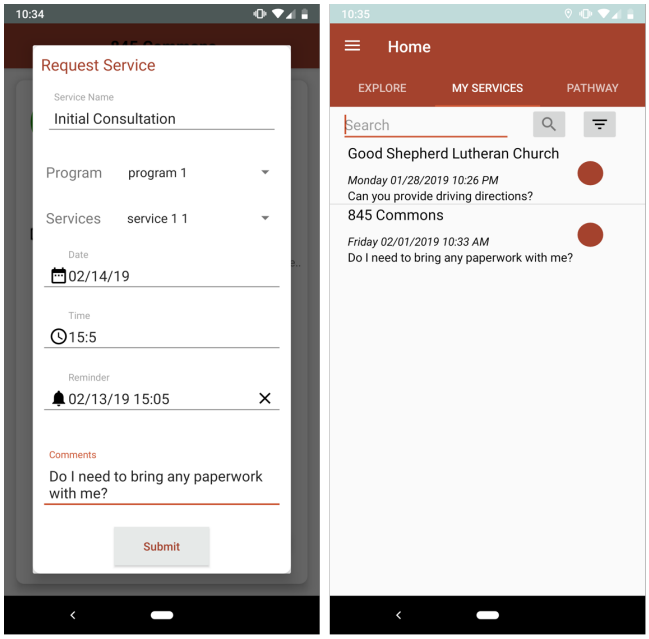
Although possible to set up our data acquisition engine to be launched in parallel using different processes and/or proxies (e.g., to parse multiple URLs in parallel), we decided against such option for two main reasons. First, we wanted to balance the use of our available bandwidth, and second, we did not want to compromise or cause any harm whatsoever to the normal operation of any of the websites visited during the data acquisition phase. Note that scrapping data may already require to break some of the terms of

[7] https://www.seleniumhq.org/
[8] https://www.guidestar.org/
[9] https://greatnonprofits.org/
[10] https://www.charitynavigator.org/



(a) (b)

Figure 3: Screenshots of our application prototype illustrating: (a) map view of human service providers color-coded by category, and (b) details about a selected provider.



(a) (b)

Figure 4: Screenshots of our application prototype showing (a) a user–friendly and easy to use dialog where the user provides inputs regarding her request to the provider, and (b) status page, where the user is able to keep track of her service requests (colors are used to differentiate between pending, cancelled, under processing, or completed requests).

use agreements used by human service providers' websites. We believe however that the benefits of crawling for data collection outweigh the potential harms.

Finally, the data acquisition engine is useful to the extent it can correctly identify the category a newly found human service provider belongs to for a large percentage of organizations based on textual information (e.g., mission statement, and available programs and services) and the category of known partnering organizations already in our repository. However, automatic categorization of organizations is a challenging problem itself. Currently, we (i) cross–reference information from GuideStar, GreatNonprofits, and Charity Navigator to assign a category to organizations already in our database, and (ii) manually annotate organizations beyond the initial list. We plan to deploy a machine learning approach to automatically perform this task as part of our classification engine module in our future work.

5 ENABLING HUMAN SERVICES 2.0

The term "Web 2.0" refers to a " *set of economic, social, and technology trends that collectively form the basis for the next generation of the Internet, a more mature, distinctive medium characterized by user participation, openness, and network effects*" [22], whereas, terms "Health 2.0" and "Medicine 2.0" have been recently used to describe the use of "Web 2.0" technologies in health care [26]. In similar spirit, we envision "Human Services 2.0", the next generation of human services, in which (i) information about service providers is openly accessible and easy to find, search and filter through, and (ii) service delivery is digital, seamless, worry–free, efficient, and traceable, maximizing positive outcomes both for service seekers and service providers alike.

To realize the potential of "Human Services 2.0", we leverage the extensive repository created by our data acquisition engine (Section 4) to open up information about human services and facilitate the digital connectivity with human service providers across conventional boundaries and silos. In our proof–of–concept prototype, a service seeker opens the application, and is prompted with a map view of human service providers nearest to her current geographic position (Figure 3a), which is provided either by allowing the application to gather her current GPS location or by dragging a pin marker to the appropriate location. A user can additionally explore human service providers in the form of a list (similar to an offline directory), or expand on any provider to obtain more information, including directions and contact details, hours of operation, programs and services provided, and eligibility criteria if applicable (Figure 3b). Filtering and search capabilities (Figure 5), and the ability to bookmark organizations (star symbol at the top right corner of Figure 3b) enable users to easily navigate among the potentially large number of service providers, whereas requesting for services become as easy as the click of a button (illustrated in Figure 4a). Tracking the status of service requests is simplified as shown in Figure 4b, whereas notifications can be set (some automatically by the app) to notify the user of changes in any of her requests (e.g., need for additional documentation). Our mobile app additionally

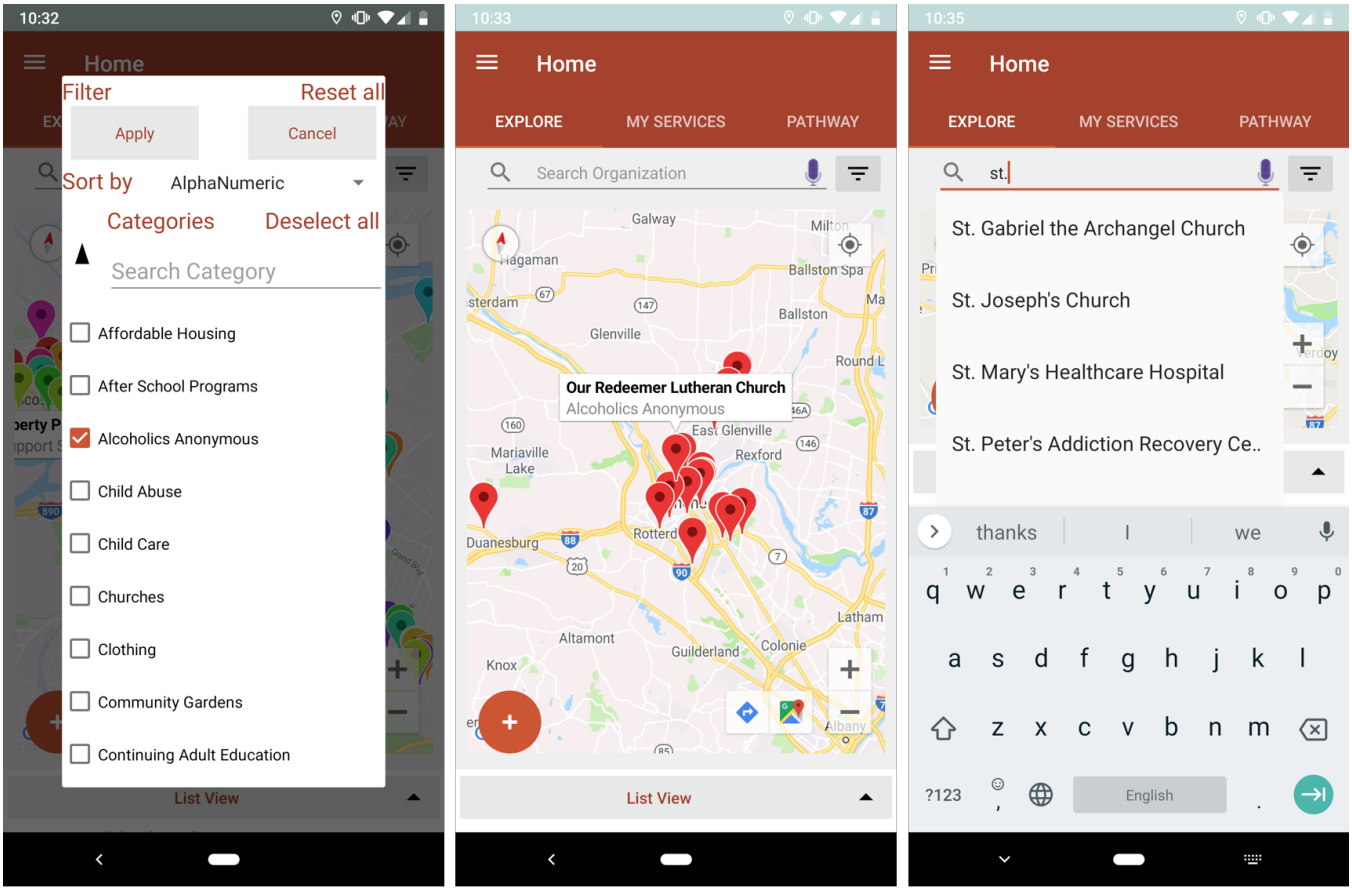


Figure 5: Screenshots of our application prototype showing (a) filtering and (c) search capabilities to narrow down one's search to a small subset of providers, as in (b) where only providers offering "Alcoholic Anonymous" services are displayed.

supports asynchronous text communication between the user and service providers.

6 CONCLUSION

The promise of *COMPASS* lies in its ability to collect and filter useful information about human service providers from a mass of semi–structured data, and democratize access to such data through a user–friendly mobile app. While still in early stages of development, *COMPASS* has generated significant interest from (i) the United Way of the Greater Capital Region[11], a non–profit organization that brings together non–profits, businesses, government and human service agencies, schools, organized labor, financial institutions, community development corporations, voluntary and neighborhood associations, and the faith community to address basic needs, education, income and health issues in the Capital Region, (ii) the Community & Public Service Program at the University at Albany, State University of New York[12], a University–wide initiative that pairs School of Social Welfare students with local non–profit and public organizations to collaboratively address issues that impact the well–being of marginalized populations, and (iii) ServeAlbany[13], a grassroots initiative with the sole goal of simplifying the process for those willing to volunteer their time and services to "*find a place to get involved and make a difference*".

Nevertheless, a number of improvements remain to be made for it to become a truly useful resource for both the general public, human service providers, and researchers. For example, we are exploring the use of additional sources to improve our data acquisition process. In pursuit of automatic categorization of newly discovered human service providers, we are exploring state–of–the–art natural language processing techniques, clustering methods, and network science principles to produce plausible labels as part of a semi–supervised or semi–automated process. With respect to the mobile app, we are working towards improving the user experience with added features such as keyword searching and improve customization such as personalized recommendations and notifications. We also plan to conduct a usability observation study to gather feedback from our target demographic on priority features

[11] https://www.unitedwaygcr.org/
[12] https://www.albany.edu/cpsp/
[13] https://servealbany.org/

as well as how to best improve the mobile app user–experience. Beyond user–evaluation, we are working towards developing quantitative evaluation metrics for the comprehensive system, including its ability to limit noise (i.e., organizations beyond the human service domain), correct classification of organizations and coverage as compared to manually curated sources such as GuideStar and GreatNonprofits.

7 ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. ECCS-1737443. We also want to thank Mr. Balaji Sankeerth Jagini for his initial contribution to the development of the data acquisition engine, Mr. Dhruvil Kotak and Bhavik Patel for their contribution towards developing our mobile app prototype, and Drs. Daphney–Stavroula Zois and Wonhyung Lee for their constructive feedback that helped improve the functionality of our mobile app prototype.

REFERENCES

[1] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. 2004. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice and Experience* 34, 8 (2004), 711–726.

[2] Judith B Bradford, Sharon Coleman, and William Cunningham. 2007. HIV System Navigation: an emerging model to improve HIV care access. *AIDS patient care and STDs* 21, S1 (2007), S–49.

[3] Deana Brown and Rebecca E Grinter. 2016. Designing for transient use: A human-in-the-loop translation platform for refugees. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 321–330.

[4] Tom Corbett and Jennifer Noyes. 2008. *Human services systems integration: A conceptual framework*. Institute for Research on Poverty.

[5] Tawanna R Dillahunt and Amelia R Malone. 2015. The promise of the sharing economy among disadvantaged communities. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2285–2294.

[6] Lynn Dombrowski, Jed R Brubaker, Sen H Hirano, Melissa Mazmanian, and Gillian R Hayes. 2013. It takes a network to get dinner: designing location-based systems to address local food needs. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 519–528.

[7] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust. 2007. Help-seeking and access to mental health care in a university student population. *Medical care* (2007), 594–601.

[8] Chris Ferguson. 2000. Shaking the conceptual foundations, too: Integrating research and technology support for the next generation of information service. *College & Research Libraries* 61, 4 (2000), 300–311.

[9] Thom Fil and Camille Ryan. 2014. *Computer and internet use in the United States*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

[10] Patricia Fredericksen and Rosanne London. 2000. Disconnect in the hollow state: The pivotal role of organizational capacity in community-based development organizations. *Public Administration Review* 60, 3 (2000), 230–239.

[11] Nancy Gourash. 1978. Help-seeking: A review of the literature. *American journal of community psychology* 6, 5 (1978), 413–423.

[12] Yeheskel Hasenfeld. 2009. *Human services as complex organizations*. Sage Publications.

[13] Allan Heydon and Marc Najork. 1999. Mercator: A scalable, extensible web crawler. *World Wide Web* 2, 4 (1999), 219–229.

[14] Marcela Horvitz-Lennon, Amy M Kilbourne, and Harold Alan Pincus. 2006. From silos to bridges: Meeting the general health care needs of adults with severe mental illnesses. *Health affairs* 25, 3 (2006), 659–669.

[15] Kenneth Kernaghan. 2005. Moving towards the virtual state: Integrating services and service channels for citizen-centred delivery. *International Review of Administrative Sciences* 71, 1 (2005), 119–131.

[16] Satya P Krishnan, Judith C Hilbert, and Dawn VanLeeuwen. 2001. Domestic violence and help-seeking behaviors among rural women: Results from a shelter-based study. *Family & Community Health* 24, 1 (2001), 28–38.

[17] Christopher A Le Dantec, Robert G Farrell, Jim E Christensen, Mark Bailey, Jason B Ellis, Wendy A Kellogg, and W Keith Edwards. 2011. Publics in practice: Ubiquitous computing at a shelter for homeless mothers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1687–1696.

[18] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3365–3376.

[19] Wonhyung Lee. 2017. Sustainability of nonprofit human service organizations in a neighborhood context. *Nonprofit Management and Leadership* 28, 1 (2017), 11–24.

[20] Ramin Mojtabai, Mark Olfson, and David Mechanic. 2002. Perceived need and help-seeking in adults with mood, anxiety, or substance use disorders. *Archives of general psychiatry* 59, 1 (2002), 77–84.

[21] Christopher Olston, Marc Najork, et al. 2010. Web crawling. *Foundations and Trends® in Information Retrieval* 4, 3 (2010), 175–246.

[22] Tim O'reilly. 2009. *What is web 2.0*. O'Reilly Media, Inc.

[23] Emmanuel Scheppers, Els Van Dongen, Jos Dekker, Jan Geertzen, and Joost Dekker. 2006. Potential barriers to the use of health services among ethnic minorities: A review. *Family practice* 23, 3 (2006), 325–348.

[24] Aaron Smith and Dana Page. 2015. US smartphone use in 2015. *Pew Research Center* 1 (2015).

[25] Michael Ungar. 2005. Pathways to resilience among children in child welfare, corrections, mental health and educational settings: Navigation and negotiation. In *Child and youth care forum*, Vol. 34. Springer, 423–444.

[26] Tom H Van De Belt, Lucien JLPG Engelen, Sivera AA Berben, and Lisette Schoonhoven. 2010. Definition of Health 2.0 and Medicine 2.0: a systematic review. *Journal of medical Internet research* 12, 2 (2010).

[27] Amy Voida, Ellie Harmon, and Ban Al-Ani. 2012. Bridging between organizations and the public: volunteer coordinators' uneasy relationship with social computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1967–1976.

Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms

Ashok Deb
USC Information Sciences Institute
Marina del Rey, CA
ashok@isi.edu

Luca Luceri*
University of Applied Sciences and Arts of Southern Switzerland, and University of Bern
Manno, Switzerland
luca.luceri@supsi.ch

Adam Badawy
USC Information Sciences Institute
Marina del Rey, CA
badawy@isi.edu

Emilio Ferrara
USC Information Sciences Institute
Marina del Rey, CA
emiliofe@usc.edu

ABSTRACT

One of the hallmarks of a free and fair society is the ability to conduct a peaceful and seamless transfer of power from one leader to another. Democratically, this is measured in a citizen population's trust in the electoral system of choosing a representative government. In view of the well documented issues of the 2016 US Presidential election, we conducted an in-depth analysis of the 2018 US Midterm elections looking specifically for voter fraud or suppression. The Midterm election occurs in the middle of a 4 year presidential term. For the 2018 midterms, 35 Senators and all the 435 seats in the House of Representatives were up for re-election, thus, every congressional district and practically every state had a federal election. In order to collect election related tweets, we analyzed Twitter during the month prior to, and the two weeks following, the November 6, 2018 election day. In a targeted analysis to detect statistical anomalies or election interference, we identified several biases that can lead to wrong conclusions. Specifically, we looked for divergence between actual voting outcomes and instances of the #ivoted hashtag on the election day. This analysis highlighted three states of concern: New York, California, and Texas. We repeated our analysis discarding malicious accounts, such as social bots. Upon further inspection and against a backdrop of collected general election-related tweets, we identified some confounding factors, such as population bias, or bot and political ideology inference, that can lead to false conclusions. We conclude by providing an in-depth discussion of the perils and challenges of using social media data to explore questions about election manipulation.

KEYWORDS

social media, political elections, data science for society

*Also with USC Information Sciences Institute, Marina del Rey, CA.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316486

ACM Reference Format:
Ashok Deb, Luca Luceri, Adam Badawy, and Emilio Ferrara. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308560.3316486

INTRODUCTION

Inherent bias of drawing conclusions from political polls stretch back to the famous headline of "Dewey Defeats Truman" in the 1948 US Presidential election [43]. Confounding factors that led to false conclusions in the 1948 election included telephone surveys which did not use robust statistical methods and an under-sampling of Truman supporters. Likewise, in 2016, many political pundits underestimated the likelihood that Donald Trump would be elected as President of the United States. The research community demonstrated a strong interest in studying social media to get a better understanding of how the 2016 events unfolded. Numerous studies concluded that social media can be a vehicle for political manipulation, citing factors such as the effect of fake news and disinformation [5, 9, 28, 29, 33, 46, 49, 51, 55], bots [7, 8, 41, 50, 53, 58, 59], polarization [3, 6], etc.

Research also suggests that social media data comes with significant biases that limit the ability to forecast offline events, e.g., the outcomes of political elections [22–26, 38], or public health issues [2, 36, 57]. Despite these well documented issues and challenges, social media are frequently relied upon and referred to as a trusted source of information to speculate about, or try to explain, offline events. One such example is the recent 2018 US Midterm elections where widespread claims of voter fraud and voter suppression appeared in the news, often based on social media reports & accounts.

In this paper, we seek to understand whether it is possible to use Twitter as a sensor to estimate the expected amount of votes generated by each state. We propose an undertaking in which we use the tweets with the hashtag #ivoted on the election day as a proxy for actual votes. At first, this seemed like a promising research direction, as tweet volumes and vote counts correlated well for 47 of the 50 states in America. We also considered if this would be a useful approach to detecting voting issues like fraud or suppression, for example by isolating statistical anomalies in

estimated and observed volumes. To get a sense of expected tweet volume, we carried out the same analysis against general keywords related to the midterm election from a month before election day through two weeks after the election. We also considered how bots may have had an influence on election manipulation narratives by measuring their activity in the social media discourse. We finally applied a political ideology inference technique and tested it to see how well it compared to an external source of polls data.

The conclusions from our analysis are complex, and this work is meant as a note of caution about the risks of using social media analysis to infer political election manipulation such as voter fraud and voter suppression.

Contributions of this work

After exploring multiple Twitter data sets and two external sources (vote counts and Gallup), we came to the following contributions:

- We explored how social media analysis carries a lot of risks involved mainly with population bias, data collection bias, lack of location-specific data, separation of bots (and organizations) from humans, information verification and fact-checking, and lastly assigning political ideology.
- We saw a significant difference in the removal of retweets in our analysis as compared with including them. However, the effect was isolated to one particular state, Texas, indicating that the sensitivity of this effect could be a factor of location.
- There is a significant difference between people's reported political ideologies using a source like Gallup versus that can be inferred on social media. It is not possible to know if this is due to limitations of political inference algorithms, confounders, population representation biases, or else.
- In the two states (NY & TX) where there was a statistically significant discrepancy between vote counts and instances of self-reported voting via #ivoted hashtags, we found only limited anecdotal evidence of tweets reporting issues of voter fraud or suppression. The divergence can possibly be explained by confounding factors, locality and selection bias, or social influence of particular candidates in those states (e.g., Alexandria Ocasio-Cortez in NY and Beto O'Rourke in TX).

BACKGROUND

The US Midterm elections were held on 6 November, 2018. They are referred to as mid-term elections because they occur in the middle of a presidential term. Senators serve for 6 years, thus, every 2 years, nearly a third of the Senators are up for re-election. The Senate is divided into 3 classes, depending on which year they were elected. Class I was elected in 2012 and are up for re-election in 2018.

For 2018, 35 Senators out of a total of 100 senators in the 115th Congress will be up for re-election. Of the 35 senators up for election, 33 are in Senate Class I and two are Senators who vacated, whereas 15 are in what is to be considered contentious races. The 33 Class I are 30 (23 Democrats (D), 5 Republicans (R), 2 Independents (I)) up for re-election and 3 Republicans (R) who are retiring. Details on the Senate seats up for re-election are in Table 1. Additionally, all 535 House of Representative seats are up for re-election every 2 years. Excluded from our analysis are the non-voting delegates for DC and the US Territories.

Table 1: US Senate Seats Up for Election in 2018

Incumbent	State	Party	Status
Tammy Baldwin	WI	D	Contested
John Barraso	WY	R	Safe
Sherrod Brown	OH	D	Contested
Maria Cantrell	WA	D	Safe
Ben Cardin	MD	D	Safe
Tom Carper	DE	D	Safe
Bob Casey	PA	D	Safe
Bob Corker	TN	R	Retiring
Ted Cruz	TX	R	Contested
Joe Donnelly	IN	D	Contested
Dianne Feinstein	CA	D	Safe
Deb Fischer	NE	R	Safe
Jeff Flake	AZ	R	Retiring
Kirsten Gillibrand	NY	D	Safe
Orrin Hatch	UT	R	Retiring
Martin Heinrich	NM	D	Safe
Heidi Heitkamp	ND	D	Contested
Dean Heller	NV	R	Contested
Mazie Hirono	HI	D	Safe
Cindy Hyde-Smith	MS	R	Contested
Tim Kaine	VA	D	Safe
Angus King	ME	I	Safe
Amy Klobuchar	MN	D	Safe
Joe Manchin	WV	D	Contested
Claire McCaskill	MO	D	Contested
Bob Menendez	NJ	D	Contested
Chris Murphy	CT	D	Safe
Bill Nelson	FL	D	Contested
Bernie Sanders	VT	I	Safe
Tina Smith	MN	D	Contested
Debbie Stabenow	MI	D	Safe
Jon Tester	MT	D	Contested
Elizabeth Warren	MA	D	Safe
Sheldon Whitehouse	RI	D	Safe
Roger Wicker	MS	R	Safe

RELATED WORK

Since the 2016 US Presidential election, there has been a big spotlight on the sovereignty of the US election system. The *Bot Disclosure and Accountability Act* of 2018[1] gave clear guidelines for what has to be disclosed by social media companies. The article *The Rise of Social Bots* [18] brought awareness to the issue of social bots in social media platforms. In [7], Bessi & Ferrara focused on social bots detection within the online discussion related to the 2016 presidential election. Other than characterizing the behavioral differences between humans and bots, there was not an in-depth analysis of any malicious intent. In this paper, we address the potential malicious activity in online political discussion along the lines of voter fraud, voter suppression, political misinformation, and then report on the biases we found.

[1] https://www.congress.gov/bill/115th-congress/senate-bill/3127/text

Voting Issues

Concerns related to voter fraud took center stage after the 2000 US Presidential election, where it was argued that the candidate with the most votes lost and the Supreme Court decided the winner [39]. Since then, a host of public debate, congressional testimony, and several new laws passed, such as the Help America Vote Act [34], which surprisingly needed to happened after the National Voter Registration Act of 1993 (NVRA).[2] The effects of the NVRA were researched by Highton and Wolfinger [32], who concluded that provisions in the NVRA would increase voter turnout by 4.7%-8.7% and that purging voter rolls of those who had not voted in the last two years would have a 2% effect. Lastly, they identified the two most vulnerable non-voting groups to be those under the age of 30 and those who moved within 2 years of an election [32].

Moreover, it has been argued that the current US voter registration has a minimal impact on registration and that there is marginal value in any updated laws [31]. Therefore, the main concern argued by both parties is voter suppression [56]. Specifically, due to recent voter identification laws, there is an increased chance of voter suppression [30]. However, in this work we seek to find instances of voter suppression from an online social media analysis. To our knowledge, this has not been done before.

Political Manipulation

Social media serve as convenient platforms for people to connect and to exchange ideas. However, social media networks like Twitter and Facebook can be used for malicious purposes [17]. Especially in the context of political discussion, there is a significant risk of mass manipulation of public opinion. Concerning the ongoing investigation of Russian meddling in the 2016 US Presidential election, Badawy et al. [4] studied political manipulation by analyzing the released Russian troll accounts on Twitter. After using label propagation to assign political ideology, they found that Conservatives retweeted Russian trolls over 30 times more than Liberals and produced 36 times more tweets. More recently, Stella et al. [52] highlighted how bots can play significant roles in targeting influential humans to manipulate online discussion thus increasing in-fighting. Especially for the spread of fake news, various studies showed how political leaning [1], age [28], and education [49] can greatly affect fake news spread, alongside with other mechanisms that leverage emotions [20, 21] and cognitive limits [44, 45]. Additionally, Dutt et al. [16] showed how foreign actors can more so than just backing one candidate or the other, often manipulate social media for the purpose of sowing discord.

Bias

Besides manipulation, other potential problems may affect data originating from online social systems. Selection bias is one such example. Concisely, this bias yields a statistically non-representative sample of the true population. A main concern outlined by Ruths and Pfeffer [48], and to a lesser degree by Malik et al. [37], is that social media samples are not representative of the whole voting population because users self-select to participate on the platform and in specific online discussions. Each social media platform has its own set of biases. Mislove et al. [40] looked specifically at the Twitter population from a location, gender, and ethnicity viewpoint. From a location perspective, they found underrepresented counties in the Mid-West and over-represented counties in highly dense urban areas [40]. Biases in the representation of gender [47], ethnicity [11], and other sources of distortions [13] can also potentially affect the inference of political ideology.

[2]https://www.justice.gov/crt/about-national-voter-registration-act

DATA

In this study, we examine different data sources to investigate and explore the risk of using social media in the context of political election manipulation.

We used Twitter as a sensor to estimate the expected amount of votes generated by each state. For this purpose, we carried out two data collections. In the first one, we gathered the tweets with the hashtag #ivoted on election day. The second collection aimed to enlarge the spectrum to a longer period of time exploiting a variety of general keywords, related to the midterm election, to collect the tweets. As a basis for comparison, we employ two external sources. The United States Election Project is used to unveil the amount of voters in each state, while Gallup to have an estimate of the political polarization both at the country level and at the state level. By means of these three data sources, we assembled five data sets (DS1-DS5), which will be analyzed in turn in the following subsections.

DS1: #ivoted Dataset. The #ivoted Dataset (DS1) gathers the tweets with the hashtag #ivoted generated on the day of the election, November 6, 2018. It should be noticed that #ivoted was promoted by Twitter and Instagram—which typically affects the hashtag spread [19, 54]—to encourage citizens to participate in the midterm elections and increase the voter turnout. We used the Python module *Twyton* to collect tweets through the Twitter Streaming API[3] during election day. The data collection time window ranged from 6 a.m. EST on November 6 (when the first polling station opened) to 1 a.m. HST on November 7 (2 hours after the last polling station closed). Overall, we collected 249,106 tweets. As a sanity check, we queried the *OSoMe API* provided by Indiana University [14]. OSoMe tracks the Twitter Decahose, a pseudo-random 10% sample of the stream, and therefore can provide an estimate of the total volume: *OSoME* contains 29.7K tweets with the #ivoted hashtag posted by 27.2K users—it is worth noting that trending topics are typically slightly over-represented in the Twitter Decahose [14, 42]—by extrapolation, this would suggest an estimated upper bound of the total volume at around 300K tweets. In addition, on election day, Twitter reported that the hashtag #ivoted was trending with over 200K tweets (cf. Fig. 1). Having collected 249K such tweets, we can conclude that we have at our disposal a nearly complete #ivoted sample dataset.

DS2 & DS3: General Midterm Dataset. In the General Midterm Dataset, we collect tweets on a broader set of keywords. Further, we consider two different time windows for the data collection. The rationale behind these choices is to evaluate the sensitivity of our study against a different, but correlated, set of data. In other words, the main purpose is to detect whether any divergence arose with

[3]Please note that we utilize the same approach for every Twitter data collection discussed in this work.

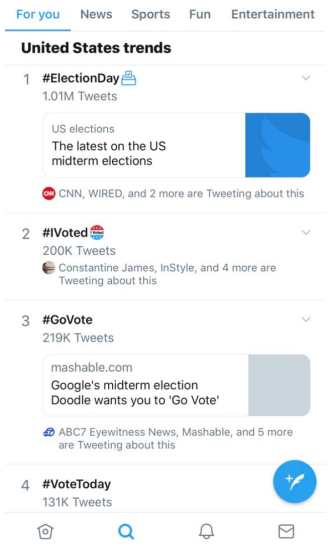


Figure 1: Screen shot of the United States trends on election day showing the #ivoted hashtag trending with 200K tweets.

the #ivoted Dataset analysis or, on the other hand, to inspect the consistency of the results in different settings.

Tweets were collected by using the following keywords as a filter: *2018midtermelections*, *2018midterms*, *elections*, *midterm*, and *midtermelections*. We distinguish two data sets according to their temporal extent. In DS2, we consider only tweets generated on the election day with exactly the same time window used for DS1. The third data set (DS3) provides a view of the political discussion from a wide-angle lens. It includes tweets from the month prior (October 6, 2018) to two weeks after (November 19, 2018) the day of the election. We kept the collection running after the election day as several races remained unresolved. As a result, DS3 consists of 2.7 million tweets, whose IDs are publicly available for download.[4]

DS4: Actual Voting Data. The first external data source used as a basis of comparison is made available by the United States Election Project. They report on their website[5] the expected voter turnout per state, along with the (official or certified) information source and other statistics about voters. The data (DS4) we use in this work was assessed on November 18, 2018, and reflects a voter turnout of 116,241,100 citizens, which is aligned with other reported counts.

DS5: Party Affiliation Data. To have an assessment of the political party affiliation across the country, we make use of an evaluation provided by Gallup, through the Gallup Daily tracking survey, a system which *continuously monitors Americans' attitudes and behaviors*.[6] The data set (DS5), collected on January 22, 2019, depicts the political leaning over a sample size of 180,106 citizens. In particular, the data shows the percentage of Democratic and Republican population in each state and over the entire country. Gallup's evaluation shows that, at the national level, there exists a democratic advantage (7%), as 45% of the population is assessed as democratic leaning while 38% is estimated as republican.

[4]https://github.com/A-Deb/midterms
[5]http://www.electproject.org/2018g
[6]https://www.gallup.com/174155/gallup-daily-tracking-methodology.aspx

Table 2: Datasets Statistics

Statistic	DS1	DS2	DS3
# of Tweets	90,763	20,450	452,288
# of Retweets	146,546	54,866	1,869,313
# of Replies	11,797	6,730	267,973
# of Authors	174,854	72,022	977,996
# of Users	178,503	77,749	997,406

Data Pre-processing

Data pre-processing involved only Twitter data sets and consisted of three main steps. First, we removed any duplicate tweet, which may have been captured by accidental duplicate queries to the Twitter API. Then, we excluded from our analysis all the tweets not written in English language. Despite the majority of the tweets were in English, and to a very lesser degree in Spanish (3,177 tweets), we identified about 59 languages in the collected tweets. Finally, we inspected tweets from other countries and removed them as they were out of the context of this study. In particular, we filtered out tweets related to the Cameroon election (October 7, 2018), to the Democratic Republic of the Congo presidential election (December 23, 2018), to the Biafra call for Independence (#biafra, #IPOB), to democracy in Kenya (#democracyKE), to the two major political parties in India (BJP and UPA), and to college midterm exams.

Overall, we count for almost 3 millions tweets distributed over the three Twitter data sets (DS1-DS3). In Table 2, we report some aggregate statistics. It should be noticed that the number of authors is lower than the number of users, which in turn also includes accounts that got a retweet (or reply) of a tweet that was not captured in our collection and, thus, they do not appear as *authors*.

METHODOLOGY

State Identification

The usage of geo-tagged tweets to assign a state to each user has been shown to not be effective, being the fraction of geo-tagged tweets around 0.5% [12]. The location of the data is of utmost importance, especially at the state and local level. However, less than 1% of the collected tweets have been geo-tagged. Nevertheless, we aim to map as many users as possible to a US state, to conduct a state by state comparison. For this purpose, we leveraged tweet metadata, which may include the self-reported user profile location. The *location* entry is a user-generated string (up to 100 characters), and it is pulled from the user profile metadata for every tweet. From this field, we first search for the two-letter capitalized state codes, followed by the full name of the state. Our analysis does not include Washington, D.C., so we have to ensure anything initially labeled *Washington* does not include any variant of *DC*. Using this string-search method, we managed to assign a state to approximately 50% of the tweets and 30% of the users. Some users had multiple states over their tweet history, thus, we only used the most common reported state. A few users often switched their location from a state name to something else: for example, one user went from

New York, NY to *Vote Blue!*—for such users, we kept the valid state location.

Bot Detection

Bot detection has received ample attention [18] and increasingly sophisticated techniques keep emerging [35]. In this study, we restrict our bot detection analysis to the use of the widely popular *Botometer*,[7] developed by Indiana University. The underpinnings of the system were first published in [15, 53] and further revised in [59]. Botometer is based on an ensemble classifier [10] fed by over 1,000 features related to the Twitter account under analysis and extracted through the Twitter API. Botometer aims to provide an indicator, namely *bot score*, that is used to classify an account either as a bot or as a human. The lower the bot score, the higher the probability that the user is not an automated and/or controlled account. In this study we use version v3 of Botometer, which brings some innovations and important detailed in [59]—e.g., the bot scores are now rescaled and not centered around 0.5 anymore.

In Figure 2, we depict the bot score distribution of the 1,131,540 distinct users in our datasets. The distribution exhibits a right skew: most of the probability mass is in the range [0, 0.2] and some peaks can be noticed around 0.3. Prior studies used the 0.5 threshold to separate humans from bots. However, according to the re-calibration introduced in the latest version of Botometer [59], along with the emergence of increasingly more sophisticated bots, we here lower the bot score threshold to 0.3 (i.e., a user is labeled as a bot if the bot score is above 0.3). This threshold corresponds to the same level of sensitivity setting of 0.5 in prior versions of Botometer (cf. Fig 5 in [59]). In both DS1 and DS3, 21.1% of the users have been classified as bots, while in DS2 the percentage achieves the 22.9% of the users. Finally, 19.5% of the 295,352 users for which a State was identified have been scored as bots.

Overall, Botometer did not return a score for 42,904 accounts, which corresponds to 3.8% of the users. To further examine this subset of users, we make use of the Twitter API. Interestingly, 99% of these accounts were suspended by Twitter, whereas the remaining 1% were protected (by privacy settings). For the users with an assigned location, only 1,033 accounts did not get a Botometer score. For those users, we assume that the accounts suspended (1,019) are bots and the private accounts (14) are humans.

Statistical Vote Comparison

Once the states have been identified and the bots detected, we compared the distribution of our various Twitter datasets (DS1, DS2, and DS3) with our control data in DS4 and DS5. To do this, we start by counting the number of tweets per state and dividing it by the total number of tweets across all states. We denote this fractional share in terms of tweets as *State Tweet Rate (STR)*, for each state i as

$$STR(i) = \frac{\text{no. tweets from State i}}{\sum_j^{50} \text{no. tweets from State j}} \quad (1)$$

For the actual voter data (DS4), we perform a similar metric to determine the *State Vote Rate (SVR)* of each state i as

[7] https://botometer.iuni.iu.edu/

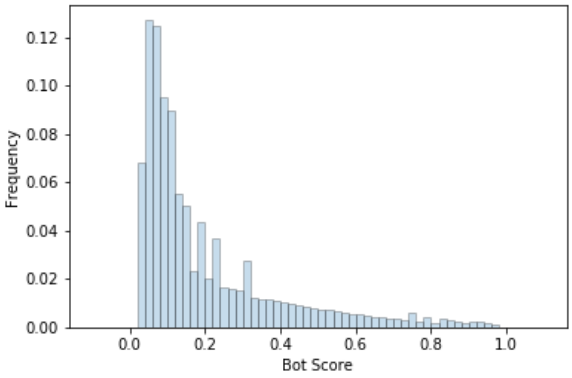


Figure 2: Bot Score Distribution

$$SVR(i) = \frac{\text{no. votes from State i}}{\sum_j^{50} \text{no. votes from State j}} \quad (2)$$

We then calculate the difference $\delta(i)$ for each state i. Here it is important to note that any positive value indicates more tweets than votes, as a percentage, and vice versa:

$$\delta(i) = STR(i) - SVR(i) \quad (3)$$

Lastly, we convert the difference into standard deviations $s(i)$ (stdevs) by dividing $\delta(i)$ by the standard deviation of all differences:

$$s(i) = \frac{\delta(i)}{\sqrt{\frac{\sum(\delta(i)-\overline{\delta})}{50}}} \quad (4)$$

being $\overline{\delta}$ the average difference over all states. We then inspect the results for any anomalous state i whose standard deviation $|s(i)| \geq 2$. States beyond two standard deviations are worth further inspection.

Political Ideology Inference

We classify users by their ideology based on the political leaning of the media outlets they share. We use lists of partisan media outlets compiled by third-party organizations, such as AllSides[8] and Media Bias/Fact Check.[9] We combine liberal and liberal-center media outlets into one list and conservative and conservative-center into another. The combined list includes 641 liberal and 398 conservative outlets. However, in order to cross reference these media URLs with the URLs in the Twitter dataset, we need to get the expanded URLs for most of the links in the dataset, since most of them are shortened. As this process is quite time-consuming, we get the top 5,000 URLs by popularity and then retrieve the long version for those. These top 5,000 URLs account for more than 254K, or more than 1/3 of all the URLs in the dataset. After cross-referencing the 5,000 long URLs with the media URLs, we observe that 32,115 tweets in the dataset contain a URL that points to one of the liberal media outlets and 25,273 tweets with a URL pointing to one of the conservative media outlets. We use a polarity rule to label Twitter users as liberal or conservative depending on the number of tweets they produce

[8] https://www.allsides.com/media-bias/media-bias-ratings
[9] https://mediabiasfactcheck.com/

with links to liberal or conservative sources. In other words, if a user has more tweets with URLs to liberal sources, he/she is labeled as liberal and vice versa. Although the overwhelming majority of users include URLs that are either liberal or conservative, we remove any user that has equal number of tweets from each side. Our final set of labeled users includes 38,920 users.

To classify the remaining accounts as liberal or conservative, we use label propagation, similar to prior work [4]. For this purpose, we construct a retweet network, containing nodes (Twitter users) with a direct link between them if one user retweet a post of another. To validate results of the label propagation algorithm, we apply stratified cross (5-fold) validation to a set of more than 38,920 seeds. We train the algorithm on 4/5 of the seed list and see how it performs on the remaining 1/5. Both precision and recall scores are around 0.89. Since we combine liberal and liberal-center into one list (same for conservatives), we can see that the algorithm is not only labeling the far liberal or conservative correctly, which is a relatively easier task, but it is performing well on the liberal/conservative center as well. Overall, we find that the liberal users population is almost three times larger the conservative counterpart (73% vs. 27%).

RESULTS

#ivoted (DS1) Statistical Analysis

There were 249,106 tweets in the #ivoted data set, of those we could map a state location for 78,162 unique authors. Once we remove the 15,856 bots (using a bot threshold score of 0.3), we have 62,306 remaining authors of tweets and retweets. After applying the method described in Statistical Vote Comparison section, we see that three states show an anomalous behavior from the remaining 47 states. Figure 4a shows how New York is 5.8 standard deviations greater than the mean difference between the #ivoted percentage and the actual voting percentage. Furthermore, both California and Texas have a stdev 2.2 greater than the mean. This would lead to believe that if there was voter suppression, it would most likely be in these three states, as they exhibit significantly more self-reported voting tweets than vote counts.

However, since our data set has both tweets and retweets, to check the sensitivity of our findings, we repeated our analysis without the retweets. Once removed, the 34,754 remaining tweets, again without bots, we noticed something interesting. Not only did Texas drop from 2.2 stdevs to 0.4 stdevs, but New York increased from 5.8 stdevs to 6.3 stdevs. This highlights the sensitivity our this type of analysis to location-specific factors such as state, and information dynamic factors such as retweet filtering. Further inspection showed that 62.2% of the tweet activity in Texas (in the #ivoted data set) was based on retweets, highlighting how this class of tweet can produce different results for some populations, and similar ones for others, since the average across the states stayed at 0 (e.g., see Figure 4b).

General Midterm (DS2&DS3) Statistical Analysis

We carried out the same analysis against the general keywords data set both on election day (DS2) and for a month before to two weeks after the election (DS3).

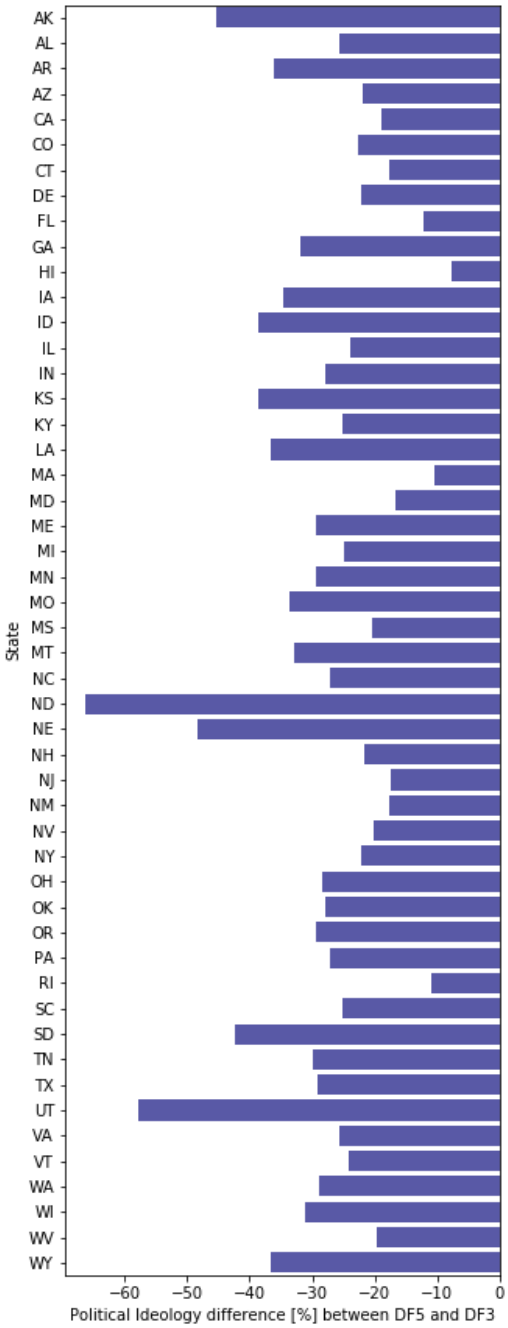


Figure 3: Political ideology difference, in terms of percentage of liberals vs. conservatives, between DS5 and DS3

In DS2, we have 72,022 users, from which we filtered out 16,859 bots (using a bot threshold of 0.3). From the remaining 55,163 authors, we were able to map a state for 26,081 users. Performing the same comparative analysis from before, we found the same anomalies in the same three states: CA (1.6 stdev), TX (2.8 stdev), and NY (5.6 stdev). Visually, this can be appreciated in Figure 4c. Expanding

the analysis to DS3, we removed 206,831 users, as classified as bots, from the set of 977,966 authors. This left us with 771,135 users from which we could identify a state for 295,705 of them. The statistical analysis revealed the same outliers also in this data set: CA (2.8 stdev), TX (3.1 stdev), and NY (4.7 stdev), as can been seen in Figure 4d.

Bot Sensitivity

Next, we investigate whether discarding malicious accounts, such as social bots, from the set of users may have affected the findings above. Table 4 shows the number (and percentage) of bots and humans per state in DS3. The list of states is sorted (in descending order) according to the percentage of bots, while the horizontal line separates the states with a bots percentage above and below the average (20.3%). Note in particular that all the three outliers (in bold) have values below the average. However, the distribution of bot prevalence per state varies greatly and it should be analyzed taking into account both the state population size and the number of Twitter users per state. Highly populated states like California, Texas, and New York, have large sheer numbers of bots but low proportional bot percentage. This should be taken into account when drawing conclusions from this analysis. On the other side, this topic opens the way to further discussions about bots association with a given state. One could make the argument that if the account was identified as a bot, there is no point to assigning it to a state. However, the fact that automated accounts declare a location in their profile can be viewed as a malicious strategy to embed in the social system thus, it should be prudently examined.

For these reasons, we repeated our analysis including social bots in the users set. Results with or without bots are substantially unchanged. In the interest of space, we do not duplicate the maps shown in Figure 4, but the same anomalies are revealed if bots are retained. It should be noticed that also for the #ivoted dataset (DS1), the percentage of bots in the three outlier states are below the average (21.0%), NY (16.0%), CA (19.4%) and TX (20.2%), respectively.

Political Ideology Analysis

Next we examine what topics talk about and how they address politically charged topics. Table 3 shows the top 10 hashtags discussed respectively by humans and bots, for both liberal and conservative ideologies. The hashtags have been colored to show the common topics between bots and humans for each political wing. The amount of overlap between bots and humans hashtags is noticeable. This is likely the reason why the removal of bots from the analyzed accounts did not have any significant impact on our outcome. To carefully interpret this table, it should be noticed that the liberal group is almost three times larger than the conservative one, as we stated in *Political Ideology* section.

Additionally, we took our political ideology labels by state and compared with DS5, the Gallup poll survey. As mentioned before, the political ideology inference assigned 73% liberal labels and 27% conservative labels to the nation at a whole. That compares with Gallup reporting of 45% to 38% for the Nation as a whole. At the state level, we ran a comparison to see the difference in our assessment of political leaning of a state versus Gallup's. For example, Alabama is 35% liberal and 50% conservative, according to

Table 3: Top 10 hashtags: liberals, conservatives, humans, bots

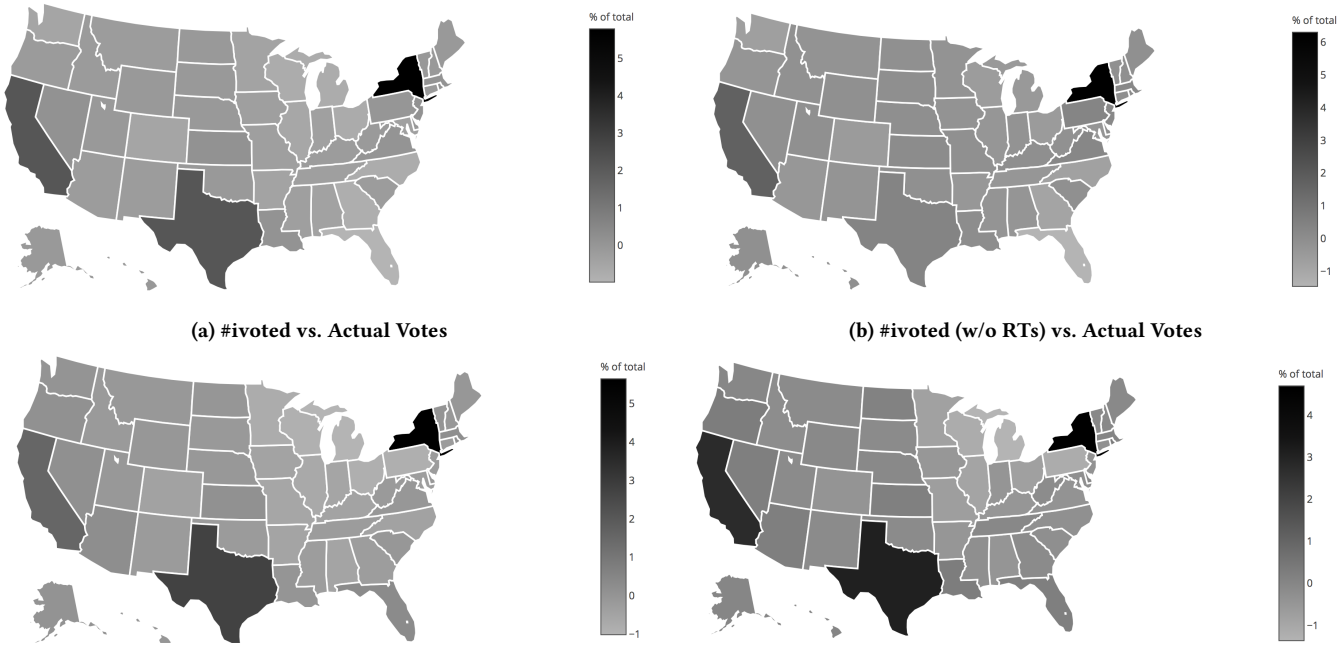
Top 10 Hashtags		
	Liberal	Conservative
Bots	#BlueWave	#BrowardCounty
	#VoteBlue	#MAGA
	#MAGA	#Broward
	#NovemberisComing	#RedWave
	#TheResistance	#VoteRedToSaveAmerica
	#Democrats	#StopTheSteal
	#Trump	#VoteRed
	#vote	#Democrats
	#Florida	#Redwavepolls
	#GOTV	#WednesdayWisdom
Humans	#NovemberisComing	#BrowardCounty
	#VoteBlue	#Broward
	#BlueWave	#MAGA
	#vote	#IranRegime
	#txlege	#Tehran
	#electionday	#StopTheSteal
	#Russia	#RedWave
	#unhackthevote	#PalmBeachCounty
	#AMJoy	#Redwavepolls
	#Trump	#Florida

Gallup, giving the state a marked Republican advantage. However, in Twitter we observed 42% Liberal and 31% Conservative user labels, which may suggest the opposite trend. Figure 3 shows the difference between the Gallup poll and our analysis. For Alabama going from a Republican advantage of 15% (Gallup) to a Democratic advantage of 11% (Twitter) would imply a shift of 26 percent points toward the liberal side. Overall, every state showed movement toward the left, as low as a few percent points and as high as over 60% difference. This corroborates the suspect that left-leaning users are over-represented in our data.

Voting Issues

New York was the state that exhibited the strongest statistical anomaly. Thus, we conducted a manual inspection reading all tweets originating from there. We found no red flags, but we isolated a few tweets of interest. The first one is in Figure 5 and it is from a user who was classified as a human and from inspection of the account shown to live in New York. The user mentions some important issues: at 11:20 am on the day of the election, they found out they are the victim of voter fraud. There is no information to suggests this was resolved in any meaningful way or if the accusation is substantiated.

A second example of potential voter issue was found after a manual inspection of the tweets in New York. The tweet thread in Figure 6 is heavily redacted, but it shows an ongoing conversation through replies and it shows multiple people presenting multiple sides. The original tweet was actually posted on 5 November, 2018 and by the time of our viewing had received a significant number of retweets. It is from this original tweet that we see a reply where the user is complaining that they can not get to the voting booth without a photo ID. User 3 then asks for the name and number of the



(a) #ivoted vs. Actual Votes

(b) #ivoted (w/o RTs) vs. Actual Votes

(c) General (election) vs. Actual Votes

(d) General (overall w/o RTs vs. Actual Votes

Figure 4: Various datasets versus Actual Votes (DS4) all without bots

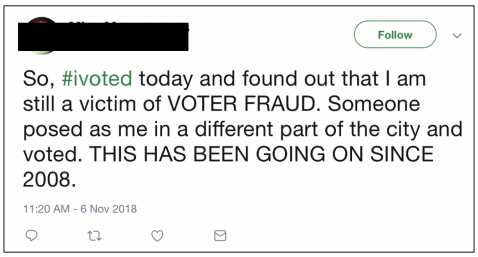


Figure 5: #ivoted tweet from New York

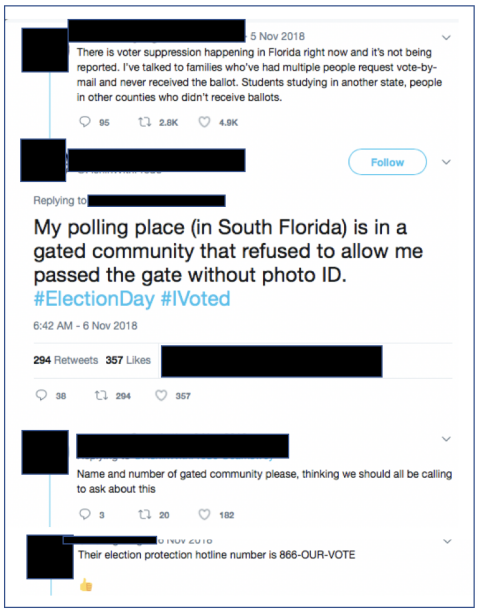


Figure 6: #ivoted tweet from Florida

community and then User 4 provides an election hotline number. This indicates that many people today are willing to speculate on Twitter, but nothing seems to indicate that they also were going to the official Department of Justice website to file a complaint.

From our inspection other tweets that are noteworthy include:

(1) "First time voter in my family registered over a month ago on DMV website online not realizing it's not automated... she could not vote. Not right."
(2) "More voter fraud in Ohio. Why is it that all the errors are always the Democrats?? Because the only way they can win is if they cheat!! This madness needs to stop."
(3) What we did see in our Twitter collection is early skepticism that there would be false claims of voter fraud. A user tweeted "a little over 24 hours from now the Racist in Chief will start Tweeting about rigged elections, voter fraud and illegal aliens voting en mass...".
(4) Shortly afterwards, many people started to retweet a user that stated "Massive voter fraud in Texas Georgia Florida and others" and also indicating that MSM (main stream media) are putting out fake polls. The Washington Post

@washingtonpost tweeted "without evidence, Trump and Sessions warn of voter fraud" which was retweeted throughout election day.

(5) There was a user who tweeted about voting machine malfunctions which mapped to a story/blog from the Atlanta Journal Constitution (https://t.com/riCGdbwQ6R) about machines being down; people left and were encouraged to come back. There was an offer for casting a paper provisional ballot, but many said they did not trust the paper ballot and wanted to vote on a machine.

DISCUSSION & RECOMMENDATIONS

Our results have highlighted the challenges of using social media in election manipulation analysis. A superficial interpretation of anomalies in online activity compared to real world data can lead to misleading or false conclusions. In our case, we wanted to determine the feasibility of using social media as a sensor to detect election manipulation such as widespread voter suppression or voter fraud. While we did not find widespread or systematic manipulation, we learned a few lessons worthy of a discussion:

- Data biases of online platforms can drastically affect the feasibility of a study. In our case, we were looking for a representative sample of actual voters who are not bots and whose political ideology and location could be known. Despite troves of data were collected and analyzed, various encountered biases could not be adjusted for.
- The second main issue is consistency in the analysis: the sensitivity to choices made when carrying out data cleaning, parameter settings of inference algorithms, etc. yield a so-called *garden of forking paths* [27]: some results can significantly vary in function of such choices (for example, location bias and the removal or retention of retweets played a role in determining whether Texas exhibited a statistical anomaly in terms of expected versus cast votes).
- Political ideologies reported by Gallup significantly vary with respect to that can be inferred on social media. We were unable to determine if this is due to limitations of the employed political inference tool, population biases, or other factors. This is an open problem in social media analysis and a necessary one to tackle before social media can be used to robustly replace polling.
- The actual voting numbers reported by official sources correlated very closely to what we inferred from our analysis on Twitter for 47 of 50 states. As such, the approach seemed promising to identify voter suppression or fraud. However, the results show a more complex picture: no evidence of fraud or suppression beyond anecdotal was found in the three anomalous states under scrutiny. Yet, we suggest that prior and during elections there should be an online social media presence for the Department of Justice to engage with people who have a potential voting issue.

CONCLUSION AND FUTURE WORK

In this work, we conducted an investigation to analyze social media during the 2018 US Midterm election. In addition to studying bots and the political ideology of users, we studied the correlation between people talking about voting and actual voter data. We then highlighted a few issues that could lead to inaccurate conclusions. In particular, removing or retaining the bots didn't change the outcome of our results. This was not the case in prior studies. However, in our case, removing retweets did make a significant difference for one state, Texas, suggesting a dependency, or bias, on location.

The challenges we faced can all be expanded upon in future work. We only mapped a state to 44.7% of DS1 and 30.2% to DS2/DS3. If we can evaluate a user timeline to better recognize what state they may be from that would enhance future location based studies. Our political ideology inference started with the labeling of 38K users leveraging any link they posted, and then labels were propagated on the retweet network. We could potentially identify the users with high centrality and evaluate their timeline for party affiliation and approach the inference problem from a different angle. We could also focus on separating not just human from bot accounts, but also human from corporate accounts. Some of the users that were classified as human could be operating as part of a collective body, that while not necessarily malicious, may insert an inorganic bias.

Ultimately, one of the goals of this work was to explore the feasibility of using social media as a sensor to detect possible election manipulation at scale: despite our initial effort did not produce the expected results, we highlighted some useful lessons that will illuminate on future endeavors to use such data for social good.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support by the Air Force Office of Scientific Research (award #FA9550-17-1-0327). L. Luceri is funded by the Swiss National Science Foundation (SNSF) via the CHIST-ERA project *UPRISE-IoT*.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] Jon-Patrick Allem, Emilio Ferrara, Sree Priyanka Uppu, Tess Boley Cruz, and Jennifer B Unger. 2017. E-cigarette surveillance with social media data: social bots, emerging topics, and trends. *JMIR public health and surveillance* (2017).
[3] Marina Azzimonti and Marcos Fernandes. 2018. *Social media networks, fake news, and polarization.* Technical Report. National Bureau of Economic Research.
[4] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *Int. Conference on Advances in Social Networks Analysis and Mining*. 258–265.
[5] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2018. Who Falls for Online Political Manipulation? *arXiv preprint arXiv:1808.03281* (2018).
[6] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *PNAS* 115, 37 (2018), 9216–9221.
[7] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).
[8] Olga Boichak, Sam Jackson, Jeff Hemsley, and Sikana Tanupabrungsun. 2018. Automated Diffusion? Bots and Their Influence During the 2016 US Presidential Election. In *International Conference on Information*. Springer, 17–26.
[9] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.
[10] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[11] Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. ePluribus: Ethnicity on Social Networks. *ICWSM* 10 (2010), 18–25.
[12] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*. 759–768.
[13] Aron Culotta. 2014. Reducing sampling bias in social media data for county health inference. In *Joint Statistical Meetings Proceedings*. 1–12.

Table 4: General Midterms DS3: bot and human population by State (sorted by percent-wise bot prevalence).

State	# of bots	# of humans
WY	97 (27.2%)	246 (68.9%)
ID	258 (23.8%)	791 (73.0%)
ND	289 (22.9%)	931 (73.9%)
AZ	1,514 (22.5%)	4,997 (74.2%)
NV	711 (22.4%)	2,377 (74.8%)
UT	420 (22.2%)	1,425 (75.4%)
DE	170 (22.1%)	575 (74.9%)
NM	325 (22.1%)	1,100 (74.7%)
NH	283 (22.1%)	968 (75.5%)
RI	402 (21.9%)	1,382 (75.3%)
WV	246 (21.7%)	854 (75.2%)
FL	4,696 (21.5%)	16,583 (75.9%)
MO	932 (21.4%)	3,336 (76.5%)
AL	697 (21.3%)	2,466 (75.4%)
TN	1,209 (21.3%)	4,369 (76.9%)
WI	808 (21.3%)	2,900 (76.4%)
MT	202 (21.0%)	730 (75.9%)
CO	1,144 (20.9%)	4,178 (76.4%)
NJ	1,311 (20.8%)	4,838 (76.8%)
MS	336 (20.6%)	1,239 (75.9%)
ME	290 (20.6%)	1,093 (77.6%)
CT	571 (20.4%)	2,141 (76.6%)
SC	769 (20.3%)	2,933 (77.5%)
OK	552 (20.2%)	2,098 (76.8%)
KS	661 (20.2%)	2,526 (77.2%)
GA	1,962 (20.2%)	7,489 (76.9%)
WA	1,561 (19.9%)	6,143 (78.2%)
NE	323 (19.9%)	1,253 (77.1%)
AK	160 (19.8%)	622 (77.1%)
HI	230 (19.8%)	895 (77.1%)
PA	1,898 (19.7%)	7,460 (77.6%)
MI	1,441 (19.6%)	5,714 (77.7%)
IA	414 (19.6%)	1,654 (78.2%)
VA	1,487 (19.6%)	5,931 (78.1%)
MA	1,372 (19.4%)	5,553 (78.4%)
NC	1,685 (19.3%)	6,872 (78.5%)
IL	1,702 (19.2%)	6,926 (78.0%)
IN	885 (19.1%)	3,593 (77.6%)
AR	199 (19.1%)	814 (78.0%)
KY	548 (19.0%)	2,270 (78.9%)
MN	866 (19.0%)	3,622 (79.6%)
OR	1,067 (18.9%)	4,416 (78.3%)
TX	**5,550 (18.7%)**	**23,448 (79.1%)**
OH	1,722 (18.6%)	7,271 (78.7%)
CA	**7,073 (18.2%)**	**30,429 (78.5%)**
VT	113 (17.9%)	505 (80.2%)
MD	887 (17.8%)	3,963 (79.7%)
NY	**4,798 (17.5%)**	**21,896 (79.9%)**
LA	708 (15.6%)	3,708 (81.7%)
SD	82 (15.4%)	439 (82.4%)

[14] Clayton A Davis, Giovanni Luca Ciampaglia, Luca Maria Aiello, Keychul Chung, Michael D Conover, Emilio Ferrara, Alessandro Flammini, Geoffrey C Fox, Xiaoming Gao, Bruno Gonçalves, et al. 2016. OSoMe: the IUNI observatory on social media. *PeerJ Computer Science* 2 (2016), e87.

[15] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*.

[16] Ritam Dutt, Ashok Deb, and Emilio Ferrara. 2018. "Senator, We Sell Ads": Analysis of the 2016 Russian Facebook Ads Campaign. In *International Conference on Intelligent Information Technologies*. Springer, 151–168.

[17] Emilio Ferrara. 2015. Manipulation and abuse on social media. *ACM SIGWEB Newsletter* Spring (2015), 4.

[18] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[19] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. 2016. Detection of promoted social media campaigns. In *Tenth International AAAI Conference on Web and Social Media*. 563–566.

[20] Emilio Ferrara and Zeyao Yang. 2015. Measuring emotional contagion in social media. *PloS one* 10, 11 (2015), e0142390.

[21] Emilio Ferrara and Zeyao Yang. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1 (2015), e26.

[22] Daniel Gayo-Avello. 2011. Don't turn social media into another'Literary Digest'poll. *Commun. ACM* 54, 10 (2011), 121–128.

[23] Daniel Gayo-Avello. 2012. " I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"–A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441* (2012).

[24] Daniel Gayo-Avello. 2012. No, you cannot predict elections with Twitter. *IEEE Internet Computing* 16, 6 (2012), 91–94.

[25] Daniel Gayo-Avello. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review* 31, 6 (2013), 649–679.

[26] Daniel Gayo Avello, Panagiotis T Metaxas, and Eni Mustafaraj. 2011. Limits of electoral predictions using Twitter. In *ICWSM*.

[27] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* (2013).

[28] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378.

[29] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5, 1 (2019), eaau4586.

[30] Zoltan Hajnal, Nazita Lajevardi, and Lindsay Nielson. 2017. Voter identification laws and the suppression of minority votes. *The Journal of Politics* 79, 2 (2017).

[31] Benjamin Highton. 2004. Voter registration and turnout in the United States. *Perspectives on Politics* 2, 3 (2004), 507–515.

[32] Benjamin Highton and Raymond E Wolfinger. 1998. Estimating the effects of the National Voter Registration Act of 1993. *Political Behavior* 20, 2 (1998), 79–104.

[33] Philip N Howard, Gillian Bolsover, Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. 2017. Junk news and bots during the US election: What were Michigan voters sharing over Twitter. *CompProp, OII, Data Memo* (2017).

[34] Brian Kim. 2003. Help America Vote Act.

[35] Sneha Kudugunta and Emilio Ferrara. 2018. Deep Neural Networks for Bot Detection. *Information Sciences* 467, October (2018), 312–322.

[36] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014).

[37] Momin M Malik, Hemank Lamba, Constantine Nakos, and Jurgen Pfeffer. 2015. Population bias in geotagged tweets. *People* 1, 3,759.710 (2015), 3–759.

[38] Panagiotis T Metaxas, Eni Mustafaraj, and Dani Gayo-Avello. 2011. How (not) to predict elections. In *3rd International Conference on Social Computing*. 165–171.

[39] Lorraine C Minnite. 2017. *The myth of voter fraud.* Cornell University Press.

[40] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM* (2011).

[41] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. 2017. Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots. *Plos One* 12, 9 (2017), e0184148.

[42] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.

[43] Frederick Mosteller and Leonard William Doob. 1949. *The pre-election polls of 1948.* Social Science Research Council.

[44] Gordon Pennycook and David G. Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* (2018).

[45] Gordon Pennycook and David G Rand. 2019. Cognitive reflection and the 2016 US presidential election. *Personality and Social Psychology Bulletin* 45, 2 (2019).

[46] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.
[47] Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On Measuring Bias in Online Information. *ACM SIGMOD Record* 46, 4 (2018), 16–21.
[48] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.
[49] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *PNAS* (2019), 201805871.
[50] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 4787.
[51] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[52] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115, 49 (2018), 12435–12440.
[53] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Int. AAAI Conference on Web and Social Media*. 280–289.
[54] Onur Varol, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2017. Early Detection of Promoted Campaigns on Social Media. *EPJ Data Science* 6, 13 (2017).
[55] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[56] Tova Wang. 2012. *The politics of voter suppression: Defending and expanding Americans' right to vote*. Cornell University Press.
[57] Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology* 57, 2 (2017).
[58] Samuel C Woolley and Douglas R Guilbeault. 2017. Computational propaganda in the United States of America: Manufacturing consensus online. *Computational Propaganda Research Project* (2017), 22.
[59] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with AI to counter social bots. *Human Behavior and Emerging Technologies* 1, 1 (2019).

Human Values and Attitudes towards Vaccination in Social Media

Kyriaki Kalimeri
Data Science Laboratory, ISI Foundation, Turin, Italy

Mariano G. Beiró
Universidad de Buenos Aires, Facultad de Ingeniería, INTECIN (CONICET), CABA, Argentina

Alessandra Urbinati
Data Science Laboratory, ISI Foundation, Turin, Italy

Andrea Bonanomi
Università Cattolica del Sacro Cuore (UNICATT), Milan, Italy

Alessandro Rosina
Università Cattolica del Sacro Cuore (UNICATT), Milan, Italy

Ciro Cattuto
Data Science Laboratory, ISI Foundation, Turin, Italy

ABSTRACT

Psychological, political, cultural, and even societal factors are entangled in the reasoning and decision-making process towards vaccination, rendering vaccine hesitancy a complex issue. Here, administering a series of surveys via a Facebook-hosted application, we study the worldviews of people that "Liked" supportive or vaccine resilient Facebook Pages. In particular, we assess differences in political viewpoints, moral values, personality traits, and general interests, finding that those sceptical about vaccination, appear to trust less the government, are less agreeable, while they are emphasising more on anti-authoritarian values. Exploring the differences in moral narratives as expressed in the linguistic descriptions of the Facebook Pages, we see that pages that defend vaccines prioritise the value of the family while the vaccine hesitancy pages are focusing on the value of freedom. Finally, creating embeddings based on the health-related likes on Facebook Pages, we explore common, latent interests of vaccine-hesitant people, showing a strong preference for natural cures. This exploratory analysis aims at exploring the potentials of a social media platform to act as a sensing tool, providing researchers and policymakers with insights drawn from the digital traces, that can help design communication campaigns that build confidence, based on the values that also appeal to the socio-moral criteria of people.

KEYWORDS

vaccine hesitancy; social media; facebook; moral foundations; political views; personality traits; network embeddings; communication campaigns

ACM Reference Format:
Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019. Human Values and Attitudes towards Vaccination in Social Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316489

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316489

1 INTRODUCTION

Over the last decades, vaccines have saved countless lives and are undoubtedly the most successful and cost-effective intervention to improve health, both in an individual and at a community level. Despite their effectiveness, vaccine hesitancy is becoming an emerging issue both in high- and low-income countries [30, 38]. According to the W.H.O. [12], to the present day, "no uniform, global metric for quantifying vaccine hesitancy currently exists". Approximately 40% of parents in the United States may delay or refuse vaccinations for their children [44].

A growing number of people use the Internet and social media to obtain information about health-related issues, including information about vaccines. Nowadays, social media platforms, such as Twitter and Facebook, are becoming increasingly more popular sources of health information [8, 29] despite their content being often subject to popularity dynamics. As a recent study shows, the consumption of vaccine-related content on Facebook is dominated by the "echo chamber" effect [40].

Reasoning on the evidence around vaccination is hardly ever the outcome of a fact-driven analysis. Vaccine hesitancy is a complex issue with numerous underlying factors influencing the decision-making process [39]. The determinants of vaccination endorsement may range from socio-demographic to psychological, emotional, and cultural factors [13, 14, 19, 43]. Political opinions, too, influence the willingness to trust advice received by governments or evidence-based medical models [28]. Since taking a stance regarding vaccination is a broader psychological and moral decision-making process, opinions are unlikely to change by appeals to reasoning or evidence alone [7]. Hence, the design of effective communication interventions requires an in-depth understanding of all the determinants of vaccine hesitancy of the specific population [2].

Psychological attributes and worldviews are reflected in a wide range of digital traces, indicatively on smartphone data [27, 48], Twitter [36], and Facebook [50]. Here, we engaged a large cohort in Italy ($N = 34,200$), using a Facebook application. Participants of this cohort were invited to complete surveys on a range of topics including psychological, moral, and political views. They allowed us to access their Likes on Facebook Pages; assuming that "liking" a Page with positive or negative attitudes towards vaccination reflect the opinion of the person on the topic. We compared the psychological and moral profile of people with a positive or negative attitude towards vaccination, presenting interesting insights on

how the two groups differ. Moreover, addressing linguistically the moral narratives provided on the description of each page, we give an insight into the values that are more emphasised on the two communities. Finally, we create a network embedding based on the Pages "Liked" by our participants to explore common interests these people may share.

The broader goal of this study is to investigate whether Facebook can be employed as a sensing tool, as a proxy to the narratives that determine vaccination endorsement. Digital data from social media can assist researchers and policymakers in understanding better the psychological profile, moral worldviews, but also the discourses underlying anti-vaccinationism on a larger scale. Such insights may inform interventions that take into consideration the cognitive, moral, psychological, and political values that are more likely to appeal to the vaccine-hesitant people.

2 DATA COLLECTION AND METHODS

2.1 Related Work and Theoretical Background

Previous research on the determinants of attitude formation towards vaccination suggests that there are broader psychological, political, cultural, or even societal factors that may contribute to negative vaccination attitudes [4, 28, 39, 49]. More specifically, vaccination scepticism has been related to unwillingness to engage with the scientific evidence[7], an alignment with alternative/complementary or holistic health [28], as well as spiritual and religious identities [7, 28], anti-authoritarian worldviews[7], conspiracy ideation [25], trust and political attitudes [49].

In this study, we operationalise the political viewpoints, moral values, personality traits, as well as cultural elements of both supporters and refuters of vaccination via self-reported questionnaires, aiming to provide a holistic view of the beliefs of the two communities.

Political Values. Our moral, psychological, and ethical values mediate relations to opinions and attitudes towards major societal issues. Inspired by the 41-item inventory proposed by Schwartz et al.[42] and the updated version of Barnea et al. [3] we included a 15-item survey to the participants of *Likeyouth* since the full version of the original inventories would be too long for an administered study on Facebook. The questions included regarding essential issues of the society, often arguments of long political debates, such for example opinions on immigration, fair treatment of individuals, trust in the Government and the European Union. The participants were asked to rate the following items, on a 5-point Likert scale, how much they agreed or not with the following phrases.

- *Q1*: It is extremely important to defend our traditional religious and moral values.
- *Q2*: It would be a good idea to privatise all of the public enterprises.
- *Q3*: People who come to live here from other countries generally make our country a better place to live.
- *Q4*: If people were treated more equally in this country, we would have many fewer problems.
- *Q5*: I trust the President of the Republic.
- *Q6*: Being rich is important to me. I want to have a lot of money and buy expensive things.
- *Q7*: I believe every person in the world should be treated in the same way. I believe that everyone should have the same opportunities in life.
- *Q8*: I trust the national government.
- *Q9*: Newer lifestyles are contributing to the breakdown of our society.
- *Q10*: The less the government gets involved with business and the economy, the better off this country will be.
- *Q11*: Foreigners that come to live in our country threaten the harmony of our society.
- *Q12*: It is extremely important to respect the freedom of individuals to be and believe whatever they want.
- *Q13*: I trust the European Union.
- *Q14*: Being successful is important to me. I like to impress people.
- *Q15*: I strongly believe that the state needs to be always aware of the threads both internal and external.

Moral Foundations. The moral values were assessed via the Moral Foundations Theory (MFT); a psychologically validated questionnaire, which focuses on the explanation of the psychological basis of morality, its origins, development, and cultural variations, and identifies the following five moral foundations [21, 22]:

- **Care/Harm**: basic concerns for the suffering of others, including virtues of caring and compassion.
- **Fairness/Cheating**: concerns about unfair treatment, inequality, and more abstract notions of justice,
- **Loyalty/Betrayal**: concerns related to obligations of group membership, such as loyalty, self-sacrifice and vigilance against betrayal,
- **Authority/Subversion**: concerns related to social order and the obligations of hierarchical relationships such as obedience, respect, and proper role fulfilment.
- **Purity/Degradation**: concerns about physical and spiritual contagion, including virtues of chastity, wholesomeness and control of desires.

These foundations collapse into two superior foundations [21]; the **individualising**, which asserts that the basic constructs of society are the individuals and hence focuses on their protection and fair treatment, and the **binding** foundation, based on the respect of leadership and traditions.

Personality Traits. The personality was assessed via the Big5 personality traits model [11, 16]; a well-established theory, which characterises personality based on the following five dimensions with universal validity [41]:

- **Openness to experiences**: inventive/ curious vs. consistent/ cautious,
- **Conscientiousness**: efficient/ organised vs. easy-going / careless,
- **Extraversion**: outgoing/ energetic vs. solitary/ reserved,
- **Agreeableness**: friendly/ compassionate vs. analytical/ detached,
- **Neuroticism**: sensitive/nervous vs. secure/ confident.

Interests & Hobbies. Additionally, to the psychologically validated questionnaires, such as the ones mentioned above, the application also contains surveys regarding interests and opinions in

general. In one of the surveys, the participants were asked to rate their interest, on a 5-point Likert scale, in the following categories: Travel, Sport, Science, Food, Culture, Nature, Society and Politics, Education, Health, Hobbies, Business, Shopping.

2.2 Data Collection

Psychological, moral, and political opinion questionnaires were administered online, on *Likeyouth*[1], a Facebook-hosted application. The Facebook platform is shown to be a valid scientific tool for administering questionnaires, with eligible population, self-selection, and behavioural biases [26].

This application acts as an innovative data-collection tool, gathering apart from the self-reported psychological and opinion assessments previously described also the "Likes" on the Facebook Pages. A consent form was obtained, from all participants, regarding a privacy agreement which they declare to accept upon registration. Once participants enter the application, they are asked to fill in their profile including basic demographic information such as gender, age, employment status, and region of residence. The application is mainly deployed in Italy; it was initially launched in March 2016 [6], while we downloaded the data on September 2018. Overall 34,200 participants (11,315 female), of average age 33 years old, entered the application. This sample is representative of the geographical distribution of the Italian population (Spearman correlation of 0.88 with the population at province level in Italy).

Among the Pages "Liked" by our participants, we searched for those who contained the stem vacc* in the Page name. These were then manually annotated according to their relevance to the topic of vaccination, avoiding irrelevant pages such for instance the musical band "The Vaccines". The relevant pages to the topic of vaccination were further annotated according to their content in supportive (hereafter, PV, for brevity) or contrary (hereafter, AV) to vaccination. In total, we result with 44 PV pages, with an average of 11,122 pages "likes", 53 AV pages, with 20,238 page "likes" on average, and 13 satyric pages (all PV), with 23,823 "likes" on average. The satirical pages were aggregated to the PV pages since they were all supportive of the vaccination.

2.3 Network Embedding Definitions

We aimed at exploring the general interests of people with conflicting views on vaccination, as expressed through their "likes" on Facebook Pages. In doing so, we propose a fully unsupervised approach based on network embeddings that exploits the intrinsic properties of the network without any a priori information about the participants.

We consider the network $G = (V, \varpi)$, where V is a set of nodes, and $\varpi : V \times V \to \mathbb{N}$ is a function defining for each pair of nodes $i, j \in V$ the weight of edge (i, j), that is w_{ij}. Since, in this study, we are interested in getting insights regarding the participants' interests, we modelled as nodes of the network our participants, while the co-occurrence of "likes" on pages are employed to create the links between the participants. Hence, if the participant i and participant j share - have "Liked" - at least one page in common, then they share a link, while the weight of the link, defined as ϖ, represents the exact number of the common pages that the two endpoints share.

The network embeddings are created employing the *node2vec* implementation [2], a well-established and efficient algorithm [17, 20]. Mapping of nodes to a feature space of lower dimension maximises the likelihood of preserving network neighbourhoods. The neighbourhood exploration is based on a second-order random walk, with two parameters, namely p and q that guide the walk. Parameter p controls the likelihood of immediately revisiting a node in the walk, while q allows differentiating the search. If $q > 1$ the walk retains a local view of the neighbourhood, otherwise the walk is more inclined to visiting nodes which are further away from the current one. Each walk, that has just traversed the link (t, i) decide the next step according to the following non-normalised transition probabilities, evaluated on each edges (i, j) leaving from i, $\pi_{ij} = \alpha_{pq}(t, j) \cdot w_{ij}$, where w_{ij} is the link's weight,

$$\alpha_{pq}(t,j) = \begin{cases} \frac{1}{p} & \text{if } d_{tj} = 0 \\ 1 & \text{if } d_{tj} = 1 \\ \frac{1}{q} & \text{if } d_{tj} = 2 \end{cases} \quad (1)$$

and d_{tj} is the shortest path distance between nodes t and j.
In this way, nodes belonging to the same neighborhood preserve a "structural" equivalence or homophily. Real networks usually exhibit a mixture of both.

3 RESULTS AND DISCUSSION

In our dataset, we have 1777 unique participants in total who expressed an interest in vaccine controversies; 1907 who followed PV pages, while 113 participants followed AV pages. Of course, a small fraction of the participants followed both PV and AV pages. In the following sections, we discuss the differences and similarities emerged from the comparison of the two populations. Note that due to the spontaneous nature of Likeyouth application, not all participants completed all the questionnaires.

Demographics. Demographic information is obtained via self-reporting once the participants enter the application. Data on employment status and residence were not sufficient for concluding the population of the two cohorts; gender information, on the other hand, was available for 1739 participants. Proportionally we have slightly more males in the PV group (62% males and 36% females), with respect to the AV group (52% males and 47% females) (p-value< 0.05 on Fisher's exact test). Such observation was reported in other scientific studies, for instance, [15, 23]. Scientists systematically try to uncover the reasons why mothers appear to be more sceptical in whether to accept a vaccine or not. It's a complex and multi-facet issue, it seems that many mothers consider their child's immune system to be "unique" [9] while their decision-making process "encompassed different factors such as social norms, past experiences, emotions, values, social network influences, and other day-to-day concerns about their child's health and well-being" [14].

Political Values. The political values survey was completed by 1232 participants who also expressed interest in vaccine controversies. Comparing the PV and AV cohorts using the Mann-Whitney U

[1] http://likeyouth.org

[2] Link to the reference implementation: https://snap.stanford.edu/node2vec/

Attribute	PV median	AV median	p-value
Q1	2.0	3.0	0.011
Q2	2.0	2.0	0.395
Q3	4.0	3.0	0.160
Q4	5.0	5.0	0.170
Q5	4.0	3.0	<0.0001
Q6	3.0	3.0	0.214
Q7	5.0	5.0	0.498
Q8	3.0	2.0	<0.0001
Q9	2.0	3.0	0.001
Q10	3.0	3.0	0.478
Q11	2.0	2.0	0.176
Q12	5.0	4.0	0.121
Q13	4.0	3.0	<0.0001
Q14	3.0	3.0	0.010
Q15	4.0	4.0	0.836

Table 1: Median values and p-values for the Mann-Whitney tests on the difference in opinions in social issues between PV and AV.

test, we found a few statistically significant differences. AV seem to trust less the governmental norms, including the President of the Republic ($Q5$, p-value < 0.001), the national government ($Q8$, p-value < 0.001), as well as the European Union ($Q13$, p-value < 0.001). These results are coherent also with the findings of Browne et al. [7], unwillingness to trust information delivered by conventional authority sources is a predictor of negative attitudes to vaccination. Further, there seems to be a propensity of the AV cohort to defend the traditional religious and moral values ($Q1$), p-value $= 0.011$, and also argue that the newer lifestyles contribute to the decline of our society ($Q9$), p-value < 0.001. This finding contrast studies carried out in the US, where people with negative views towards vaccination seem to value more the notion of freedom, concerning personal expression, and religious beliefs, to the vaccine supporters [1]. Table 1, reports the differences in core political values between the two cohorts.

Moral Values. Overall, 128 participants interested in vaccine controversy page filled in the moral foundations questionnaire. To compare the two cohorts, i.e. the population that supports vaccination practices, PV, to the one that is more sceptical about it, AV, we employed the non-parametric Mann-Whitney U test [33]. Due to the limited size of the cohort, none of the results is statistically significant.

Personality Traits. Only a small fraction of the participants completed the personality traits questionnaire, 116 participants in total. Due to the small sample size, the only statistically significant result was the tendency of the PV participants to score higher in the agreeableness trait (p-value< 0.05). Table 3 summarised the findings of all the five traits. Low agreeableness scores have been previously related to vaccine hesitancy [31].

Interests and Hobbies. Self-reported information quantifying the broader interest in general topics, for instance "science", was available for 153 participants. Table 4 reports the Mann-Whitney U test

Attribute	PV median	AV median	p-value
Care	19.0	21.0	0.09
Fairness	21.0	22.0	0.42
Loyalty	15.0	18.0	0.899
Authority	14.0	13.0	0.667
Purity	14.0	14.0	0.390
Individualism	40.0	43.0	0.213
Social binding	43.0	39.0	0.921

Table 2: Median values and p-values for the Mann-Whitney tests on the difference in moral values between PV and AV participants.

Attribute	PV median	AV median	p-value
Extraversion	8.0	8.0	0.793
Agreeableness	10.0	8.0	0.025
Conscientiousness	9.0	11.0	0.157
Openness	8.0	9.0	0.315
Neuroticism	9.5	9.5	0.924

Table 3: Median values and p-values for the Mann-Whitney tests on the difference in personality traits between PV and AV participants.

Attribute	PV median	AV median	p-value
Travel	3.0	4.0	0.561
Sport	3.0	4.0	0.027
Science	4.0	3.0	0.356
Food	4.0	4.0	0.522
Culture	4.0	5.0	0.744
Nature	4.0	4.0	0.426
Society & Politics	4	4	0.976
Education	4	4	0.749
Health	3.0	4.0	0.023
Hobbies	3.0	4.0	0.049
Business	4	4	0.473
Shopping	2	2	0.423

Table 4: Median values and p-values for the Mann-Whitney U test on the difference in topics of interest between PV and AV participants.

on the two cohorts pointed out that AV participants are more into hobbies in general, but interestingly, are significantly more interested in sports and health-related topics. Such indication points out that since these individuals are more interested in health, will actively look for information about vaccination; the challenge lies in the documentation they will be more likely to receive [40, 47] but also their predisposition in interpreting what they read [5, 8, 28].

3.1 Moral Values in Facebook Pages

Having assessed the emerging differences in the political, moral, and psychological worldviews of individuals with demonstrated positive or negative attitudes towards vaccination, we place the

focal point on the understanding of the content presented in the PV and AV pages from a linguistic perspective.

To this extent, we present an exploratory visualisation of the words contained in the descriptions of the pages associated with each group (PV and AV), as a wordcloud. A wordcloud is a graphical representation of a "cloud" of words in which each word is sized according to its multiplicity in the document. Figure 1 depicts the most common words that appear in the description of the PV and AV groups of pages [3].

In a glance, we observe that PV pages emphasise the concepts of children and parents, join and love, misinformation, unite. On the other hand, AV pages focus on the idea of free choice, purpose, obligation and power. This is in line with the study of Amin et al. [1] who stated that vaccine-hesitant individuals are emphasising more the moral value of "liberty", the so-called sixth moral foundation. The above items are expressing a similar notion to the ones included in the "liberty foundation" scale employed to quantify this scale [24]. "Scientific" and "book", also appears in the AV discourses, indicating that AV community is not unwilling to engage with scientific evidence, but perhaps there is a biased interpretation of experimental outcomes [7]. Diving further into the moral rhetoric of the PV and AV page descriptions, we employed the dictionary proposed by Graham et al. [18] That relates specific words to the five predefined dimensions defined at the MFT. We found that PV pages contain more words associated with the moral dimension of care and authority while AV pages focus more on the value of loyalty. A limitation of the dictionary though is that the words related to each trait are few and some of them too formal for the everyday communication used in social media.

3.2 Facebook Page Embeddings

Network embeddings are an expanding field [45], increasingly employed in a variety of tasks, such as network structure identity and community detection [34, 37, 46, 46], but also topics of social interest such as classification of demographics and other behavioural attributes [35, 51], to social interactions and behaviour prediction [10]. Here, we propose an embedding based on the co-occurrence of pages people "Liked" on Facebook. Such vision of the network allows us to explore common interests that two individuals may have since when two nodes are physically closer in the network, the more Facebook pages they share (e.g. have commonly "Liked"). Initially, we employed this technique to pages related to "health", since from the previous analysis it emerged as a domain with a significant differentiation in the PV and AV cohorts.

We perform parameter tuning with grid search of p and q in the set {4,2,1,0.8,0.1}, always keeping $q < 1$ since we are interested in the community structure of the network and the notion of homophily. We visualise the embedding in two dimensions employing the t-SNE algorithm [32]; Figure 2 depicts the network generated by the likes in the pages related to health topics. Zooming in the area with the highest concentration of AV participants, we retrieve the Facebook pages commonly "Liked" by these participants (Table 5). We observe that many of the emerging pages are related to food and nutrition. In particular, participants in this area "Like" pages about complementary and alternative medicine like for instance, "Apoteca Natura" (Natural Pharmacy), "Cure Naturali" (Natural Cures), and "Omnama" (spiritual consultancy page), as well as medical doctors who support anti-vaccination attitudes ("Dr. Roberto Gava"). From the emerging topics, we see that these interests are in line with the claims that negative attitudes to vaccination are related to the preferences of alternative medicine, as well as the endorsement of spirituality [7]. This alternative view of the network, allows us to explore interests that people with shared views on a specific topic may share in a data-driven way.

[3]For the translation we have used the library Python Googletrans 2.4.0.

Facebook Page	Total Likes
Psychology (Psicologia)	19
The delicious good eating (Il Goloso Mangiar Sano)	17
VaccinYES (VaccinarSI)	16
Omnama	12
Self-weaning (Autosvezzamento)	12
Natural Cures (Cure Naturali)	11
You know health and nutrition (Lo Sai Salute e Alimentazione)	10
San Donato Hospital Group (Gruppo Ospedaliero San Donato)	9
Delicious Magazines (Buonissimo Magazine)	9
FruitsWeb (FruttaWeb)	9
wellMe.it	8
Rita Levi Montalcini	8
Dr. Roberto Gava	8
Natural Pharmacy (Apoteca Natura)	6

Table 5: The 15 most liked Facebook Pages by participants in the area with the highest concentration of AV supporters, see Figure 2, red box.

4 LIMITATIONS

There are several limitations to our work. First and foremost, population in our cohort that has filled in questionnaires, especially the ones about morals and personality is small, affecting the statistical significance of our results substantially. Also, many people interested in vaccine controversies are not necessarily following a Facebook page on the argument; hence, do not result in our analysis.

5 CONCLUSIONS AND FUTURE WORK

Decision-making processes are complex; multiple factors contribute to how people rationalise, form opinions, and finally take actions. When it comes to critical societal issues such as vaccine hesitancy, it is essential to understand the driver of this process, aiming at bridging the gap between the two sides. The determinants of attitude formation towards vaccination may range from psychological, moral, cultural, or even societal. [4, 28, 49].

In this study, we analysed the psychological, moral, and political views of a large cohort in Italy, employing surveys administered on a Facebook-hosted application. We assumed that "liking" a Page that expresses a positive or negative attitudes towards vaccination, the participant shares the same views on the topic.

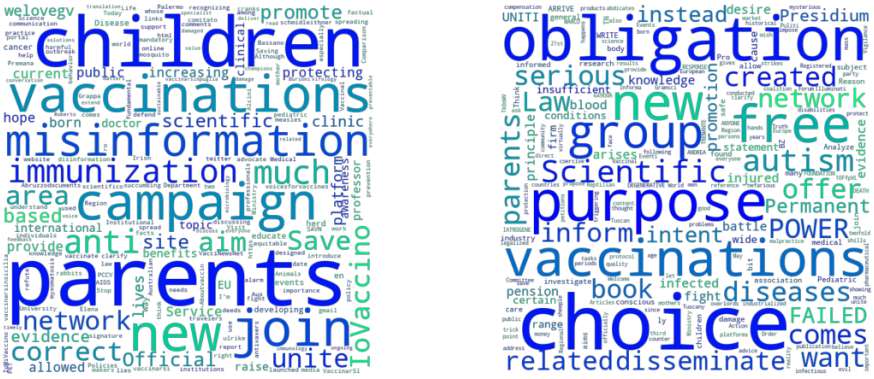


Figure 1: Wordcloud showing the most used words by PV pages (left) and AV pages (right).

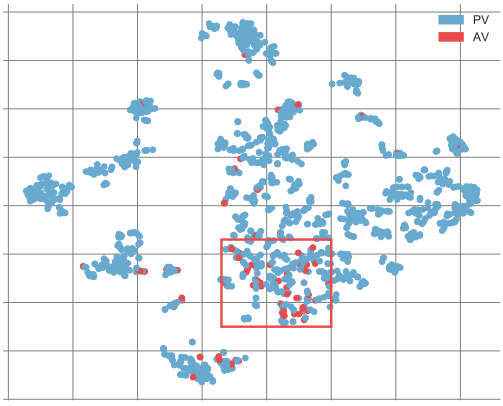


Figure 2: Health co-occurence network embedding with t-SNE visualization. Embedding parameters: dimension=50, walk length = 5, number of walks = 50, p=2,q=0.8

Analysing people's political views, we found those who are sceptical towards vaccination trust less the legal norms, both in a national and at a European level. Contrary to what is reported by a study carried in the US [1], vaccine hesitancy is related to a tendency to argue that newer lifestyles contribute to the decline of the society while feeling the urge to defend more the traditional religious and moral values. The personality determinants of vaccine hesitancy relate to low agreeableness and high openness to experience traits. Regarding general interests, vaccine-hesitant individuals appeared to be more interested in topics like sports, hobbies and health.

Moving from surveys to digital data, we focused on linguistic usage employed to express the moral narratives, in the description of the Facebook pages, of both communities (AV and PV). In line with the moral values expressed above, the narratives of supporters of vaccination and pro-vaccine campaigns emphasise more in notions of care and protection, while the opposite side emphasises more on the value of freedom of choice, science, power, and law, which relate to authority. Finally, from the network embedding based on the Pages "Liked" related to "health", we find pages related to alternative medicine and spirituality. This alternative view of the network, allows us to explore interests that people with shared views on a specific topic may share in a data-driven way.

As Debe et al. [13] suggest, due to the multi-facet nature of vaccine hesitancy there is still no effective communication strategy. Here, we propose an approach where a social media platform is employed as an alternative tool to help researchers and policy-makers better understand the psychological and moral views of vaccine-hesitant individuals.

Future steps include, first of all, the expansion of the cohort. Secondly, in this study we only assessed the linguistic content of the page descriptions; we plan to expand this task in the actual content of the pages since this will provide a more unobstructed view on the arguments made by both sides.

In conclusion, insights from social media data may inform interventions that take into consideration the cognitive, moral, psychological, and political values of people hesitant towards vaccination, building general confidence by appealing to their values.

ACKNOWLEDGMENTS

The authors acknowledge support from the "Lagrange Project" of the ISI Foundation funded by the Fondazione CRT. They would like to thank all the participants in the cohort.

REFERENCES

[1] Avnika B. Amin, Robert A. Bednarczyk, Cara E. Ray, Kala J. Melchiori, Jesse Graham, Jeffrey R. Huntsinger, and Saad B. Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour* 1 (2017), 873–880. Issue 12. https://doi.org/10.1038/s41562-017-0256-5

[2] Katie Attwell, Mark C Navin, Pier Luigi Lopalco, Christine Jestin, Sabine Reiter, and Saad B Omer. 2018. Recent vaccine mandates in the United States, Europe

and Australia: a comparative study. *Vaccine* 36, 48 (2018), 7377–7384.
[3] Marina F Barnea and Shalom H Schwartz. 1998. Values and voting. *Political psychology* 19, 1 (1998), 17–40.
[4] Sandra J Bean. 2011. Emerging and continuing trends in vaccine opposition website content. *Vaccine* 29, 10 (2011), 1874–1880.
[5] Cornelia Betsch, Noel T Brewer, Pauline Brocard, Patrick Davies, Wolfgang Gaissmaier, Niels Haase, Julie Leask, Frank Renkewitz, Britta Renner, Valerie F Reyna, et al. 2012. Opportunities and challenges of Web 2.0 for vaccination decisions. *Vaccine* 30, 25 (2012), 3727–3733.
[6] Andrea Bonanomi, Alessandro Rosina, Ciro Cattuto, and Kyriaki Kalimeri. 2017. Understanding Youth Unemployment in Italy via Social Media Data.
[7] Matthew Browne, Patricia Thomson, Matthew Justus Rockloff, and Gordon Pennycook. 2015. Going against the herd: psychological and cultural factors underlying the "vaccination confidence gap". *PLoS One* 10, 9 (2015), e0132562.
[8] Gabriela Capurro, Josh Greenberg, Eve Dubé, and Michelle Driedger. 2018. Measles, Moral Regulation and the Social Construction of Risk: Media Narratives of "Anti-Vaxxers" and the 2015 Disneyland Outbreak. *Canadian Journal of Sociology* 43, 1 (2018), 25–48.
[9] JA Cassell, M Leach, MS Poltorak, CH Mercer, A Iversen, and JR Fairhead. 2006. Is the cultural context of MMR rejection a key to an effective public health discourse? *Public health* 120, 9 (2006), 783–794.
[10] Yoon-Sik Cho, Greg Ver Steeg, Emilio Ferrara, and Aram Galstyan. 2016. Latent space model for multi-modal social data. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 447–458.
[11] Paul T Costa Jr and Robert R McCrae. 1992. The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders* 6, 4 (1992), 343.
[12] Gretchen J Domek, Sean T O'Leary, Sheana Bull, Michael Bronsert, Ingrid L Contreras-Roldan, Guillermo Antonio Bolaños Ventura, Allison Kempe, and Edwin J Asturias. 2018. Measuring vaccine hesitancy: Field testing the WHO SAGE Working Group on Vaccine Hesitancy survey tool in Guatemala. *Vaccine* 36, 35 (2018), 5273–5281.
[13] Eve Dubé, Dominique Gagnon, Noni E MacDonald, et al. 2015. Strategies intended to address vaccine hesitancy: Review of published reviews. *Vaccine* 33, 34 (2015), 4191–4203.
[14] Eve Dubé, Maryline Vivion, Chantal Sauvageau, Arnaud Gagneur, Raymonde Gagnon, and Maryse Guay. 2016. "Nature Does Things Well, Why Should We Interfere?" Vaccine Hesitancy Among Mothers. *Qualitative Health Research* 26, 3 (2016), 411–425.
[15] Melissa B Gilkey, Annie-Laurie McRee, and Noel T Brewer. 2013. Forgone vaccination during childhood and adolescence: findings of a statewide survey of parents. *Preventive medicine* 56, 3-4 (2013), 202–206.
[16] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
[17] Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151 (2018), 78–94.
[18] Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96, 5 (2009), 1029.
[19] Joshua Greenberg, Eve Dubé, and Michelle Driedger. 2017. Vaccine hesitancy: in search of the risk communication comfort zone. *PLoS currents* 9 (2017).
[20] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.
[21] Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20, 1 (2007), 98–116.
[22] Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133, 4 (2004), 55–66.
[23] EKOS Research Associates Inc. 2018. Survey for the Development of the Childhood Vaccination Campaign. *Vaccine* (2018).
[24] Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. 2012. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one* 7, 8 (2012), e42366.
[25] Daniel Jolley and Karen M Douglas. 2014. The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one* 9, 2 (2014), e89177.
[26] Kyriaki Kalimeri, Mariano G Beiró, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019. Evaluation of Biases in Self-reported Demographic and Psychometric Information: Traditional versus Facebook-based Surveys. *arXiv preprint arXiv:1901.07876* (2019).
[27] Kyriaki Kalimeri, Mariano G Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior* 92 (2019), 428–445.
[28] Anna Kata. 2010. A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine* 28, 7 (2010), 1709–1716.
[29] Allison Kennedy, Katherine LaVail, Glen Nowak, Michelle Basket, and Sarah Landry. 2011. Confidence about vaccines in the United States: understanding parents' perceptions. *Health affairs* 30, 6 (2011), 1151–1159.
[30] Heidi J Larson, Louis Z Cooper, Juhani Eskola, Samuel L Katz, and Scott Ratzan. 2011. Addressing the vaccine confidence gap. *The Lancet* 378, 9790 (2011), 526–535.
[31] Carol HJ Lee, Isabelle M Duck, and Chris G Sibley. 2017. Personality and demographic correlates of New Zealanders' confidence in the safety of childhood vaccinations. *Vaccine* 35, 45 (2017), 6089–6095.
[32] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[33] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
[34] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 701–710.
[35] Bryan Perozzi and Steven Skiena. 2015. Exact Age Prediction in Social Networks. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, 91–92. https://doi.org/10.1145/2740908.2742765
[36] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 180–185.
[37] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 385–394.
[38] Daniel A Salmon, Matthew Z Dudley, Jason M Glanz, and Saad B Omer. 2015. Vaccine hesitancy: causes, consequences, and a call to action. *Vaccine* 33 (2015), D66–D71.
[39] Philipp Schmid, Dorothee Rauber, Cornelia Betsch, Gianni Lidolt, and Marie-Luisa Denker. 2017. Barriers of influenza vaccination intention and behavior–a systematic review of influenza vaccine hesitancy, 2005–2016. *PloS one* 12, 1 (2017), e0170550.
[40] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. 2018. Polarization of the vaccination debate on Facebook. *Vaccine* 36, 25 (2018), 3606–3612.
[41] David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. 2007. The geographic distribution of Big Five personality traits patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology* 38, 2 (2007), 173–212.
[42] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*. Vol. 25. Elsevier, 1–65.
[43] Gilla K Shapiro, Ovidiu Tatar, Eve Dube, Rhonda Amsel, Barbel Knauper, Anila Naz, Samara Perez, and Zeev Rosberger. 2018. The vaccine hesitancy scale: Psychometric properties and validation. *Vaccine* 36, 5 (2018), 660–667.
[44] Philip J Smith, Sharon G Humiston, Edgar K Marcuse, Zhen Zhao, Christina G Dorell, Cynthia Howes, and Beth Hibbs. 2011. Parental delay or refusal of vaccine doses, childhood vaccination coverage at 24 months of age, and the Health Belief Model. *Public health reports* 126, 2_suppl (2011), 135–146.
[45] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.
[46] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1225–1234.
[47] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. 2018. The spread of medical fake news in social media–The pilot quantitative study. *Health Policy and Technology* (2018).
[48] Runhua Xu, Remo M. Frey, Elgar Fleisch, and Alexander Ilic. 2016. Understanding the Impact of Personality Traits on Mobile App Adoption – Insights from a Large-Scale Field Study". *Computers in Human Behavior* 62 (sep 2016), 244–256.
[49] Ohid Yaqub, Sophie Castle-Clarke, Nick Sevdalis, and Joanna Chataway. 2014. Attitudes to vaccination: a critical review. *Social science & medicine* 112 (2014), 1–11.
[50] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.
[51] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2016. Homophily, structure, and content augmented network representation learning. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 609–618.

Cross-lingual Perspectives about Crisis-Related Conversations on Twitter

Johnny Torres
Escuela Superior Politécnica del Litoral (ESPOL)
Guayaquil, Ecuador
jomatorr@espol.edu.ec

Carmen Vaca
Escuela Superior Politécnica del Litoral (ESPOL)
Guayaquil, Ecuador
cvaca@fiec.espol.edu.ec

ABSTRACT

The role of social networks during natural disasters is becoming crucial to share relevant information and coordinate relief actions. With the reach of the social networks, any user around the world has the possibility of interact in crisis-events as these unfold. A large part of the information posted during a disaster uses the native language where the disaster occurred. However, there are also users from other parts of the world who can comment about the event, often in another language. In this work, we conducted a study of crisis-related tweets about the earthquake that occurred in Ecuador in April 2016. To that end, we introduce a new annotated dataset in both Spanish and English languages with approximately 8K tweets; half of them belong to conversations. We evaluate several neural architectures to identify crisis-related tweets in a multi-lingual setting, and we found that deep contextual multi-lingual embeddings outperform other strong baseline models. We then explore the type of conversations that occur from the perspective of different languages. The results show that certain types of conversations occur more in the native language and others in a foreign language. Conversations from foreign countries seek to gather situation awareness and give emotional support, while in the affected country the conversations aim mainly to humanitarian aid.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Information systems** → *Collaborative and social computing systems and tools*; *Social networks.*

KEYWORDS

Social Computing; NLP; Neural Networks

ACM Reference Format:
Johnny Torres and Carmen Vaca. 2019. Cross-lingual Perspectives about Crisis-Related Conversations on Twitter. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316799

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316799

Table 1: A conversation about the Earthquake in Ecuador initiated by an organization's Twitter account.

User	Tweet	In reply to
BBCBreaking	Ecuador declares state of emergency in six provinces after powerful earthquake kills at least people	
RNexists	is the current count!	BBCBreaking
MaJoJovi	The current count is , I'm from Ecuador. Please pray for us.	Rnexists
1SHeRA1	I am and will, my dear. I hope we can help in some concrete ways as well.	MaJoJovi
MaJoJovi	we need all the help we can get. Your prayers helps too. Thank you.	1SHeRA1

1 INTRODUCTION

During crisis events like earthquakes, user-generated data on social media can provide valuable information to humanitarian organizations such as the United Nations, Red Cross, and also activists working in the relief efforts. Social media leverage the power of the crowds to provide awareness of the situation often faster than traditional media, allows to respond quickly to the urgent needs of affected people, assess damages in the buildings and infrastructure, identify medical emergencies, or coordinate relief actions [8, 11]. With the worldwide reach of social media, users from different countries and languages can react and interact in any crisis event. In this paper, we aim to identify the crisis-related tweets in a multi-lingual scenario and characterize them in the context of conversations. To that end, we introduce a new annotated dataset about the Earthquake occurred in Ecuador on April 16, 2016. The corpus contains 8360 tweets annotated for English and Spanish language. The table 1 shows an example of a conversation between several users on Twitter. Based on the metadata (*in reply to* field) provided by Twitter, we can collect all the tweets that belong to conversations.

For the annotation task, we leverage on taxonomy proposed by Imran et al. [13] , in which a *tweet* can belong to one of several categories such as: *statistics about affected people*, *emotional support*, or *helping* through donations, goods, or volunteers. To categorize tweets, we have to deal with some issues associated with social media data that include: a) associating a tweet to a category can be difficult due to ambiguity, even human annotators may differ in their judgment about whether or not a tweet belongs to a specific category, b) the noisy nature of the tweets, as well as the

idiomatic phrases, can make it difficult to train models and infer across languages.

Despite recent progress in natural language processing (NLP), the semantic interpretation of noisy short-texts remains a hard problem. A multi-lingual scenario difficult even more the interpretation task. To tackle these issues, we evaluate a neural architecture to identify crisis-related tweets across multiple languages, specifically in the Ecuadorian earthquake study case. The contribution of our work is as follow. a) we introduce an annotated corpus of crisis-related tweets for Spanish and English language, b) we evaluate deep contextual neural architectures for the multi-lingual classification task at hand, and c) we characterize the conversations from locals and foreigners about the study case earthquake. We made available the dataset and the code to the community[1].

The rest of the paper is structured as follows: we begin with a discussion of related work in section 2, followed the description of the dataset in section 3. Then, we describe the taxonomy and the annotation process in section 4. Next, we present the proposed cross-lingual classification model in section 5; and, we discuss the results and the limitations in section 6. Finally, we outline the conclusions and future work in section 8.

2 RELATED WORK

Previous works analyze the usefulness of large stream data from social media during crisis events. Immediately after a crisis or natural disaster has occurred, people use social media platforms to report the situation in the affected places, look for useful information, and request/offer help [8, 11].

It can be crucial during natural disasters to gain insight into the situation as it unfolds for the relief efforts by organizations and activists [1, 24, 27]. Several systems have been implemented to classify, extract, and summarize crisis-relevant information from social media [11].

Imran et al. [12] implemented the AIDR platform to collect and classify data streams during crises on Twitter. The AIDR platform has been instrumental in creating public datasets[2] to advance the research in the area of crisis informatics. Although, there is a large number of crisis events including earthquakes, floods, hurricanes, and cyclones; to the best of our knowledge the Ecuadorian earthquake is missing, and this work contributes annotated data for this event.

Several learning methods have been proposed to classify and categorize crisis-related data. Verma et al. [29] used Naive Bayes and MaxEnt classifiers to learn situational awareness tweets from several crises. Cameron et al. [7] proposed a framework for emergency awareness using an SVM classifier to identify useful tweets during natural disasters. More recently, Imran et al. [11] uses a traditional approach that requires manually engineered features like cue words and TF-IDF vectors for learning representation and training the model.

Traditional classification approaches have several shortcomings due to the discrete word representations and the dependency on the trained data for a specific event and language, so they have poor performance classifying data for new events even more for new languages (*out-of-event data*). Imran et al. [14] tackle this problem and show the performance of some non-neural network classifiers trained on labeled data from past crisis events. Recent approaches use neural architectures to deal with the issue the *out-of-event data*, specifically Convolutional Neural Networks (CNN) [6, 20] and semi-supervised learning [3, 4].

Social media reach worldwide audiences; therefore learning methods have to deal with a myriad of languages for user-generated data during crisis events. To tackle this problem, Khare et al. [15] evaluated several scenarios of learning methods with cross-lingual data. They use 30 cross-lingual datasets of crisis events where the model is trained on one language and tested in another language. Their approach uses an SVM classifier with several semantic features in addition to the tweets' text. We leverage the experimental setup proposed by Khare et al. [15] and evaluate an end-to-end learning based multi-lingual deep contextual embeddings and neural architectures, focusing only on the Ecuadorian Earthquake.

Table 2: Statistics of the tweets by language. The first section refers to all tweets in the dataset. The second section refers to the tweets that belong to conversations with at least one reply.

	Spanish	English	Other	Total
Tweets	93,405	38,533	20,331	152,269
Users	50,758	25,880	10,387	87,025
Avg. tweets	1.84	1.49	1.96	1.75
Conversations	4,632	1,092	377	6,101
Replies	50,747	17,989	9,506	78,242
Tweets	55,379	19,081	9,883	84,343
Avg. replies	11.96	17.47	26.21	13.82
Avg. users	8.34	13.65	16.36	9.79

3 DATASET

The dataset presented in this paper contains tweets provided by AIDR Research Lab[3], a platform that collects crisis-related tweets [12]. The initial dataset contained tweets collected using the Twitter Streaming API for the Ecuadorian Earthquake in April 2016. The Streaming API allows anyone to retrieve at most a 1% sample of all the Twitter's data filtered by some parameters [19]. We then augmented the dataset by retrieving the entire conversation tree for the tweets in the initial dataset. To collect the complete conversations, we crawl the parent tweets using the field *[in reply to status id]* in the tweets' metadata. From the parent tweets (i.e., the tweets that initiate a conversation), we crawled all the child tweets following the procedure in [26].

The Table 2 summarizes some statistics about the dataset. The first section shows all the tweets in the dataset split by language: Spanish, English, and other languages. We noticed that for *other* languages, the number of the average number of tweets is higher than English or Spanish.

In the second section, we filter out isolated tweets and use only tweets that form part of conversations with at least one reply. The

[1] https://github.com/johnnytorres/crisis_conv_crosslingual
[2] http://crisisnlp.qcri.org/
[3] http://aidr.qcri.org/

conversations row refers to the number of conversations by language. We identify each conversation by its initial tweet. Then, we have the number of *replies*, in total 55%, of the tweets in the dataset belong to a conversation. Although the average of the number of the number of replies and users vary across languages, the median resulted similar for all languages (num replies = 3 and num users = 2) due to the constraint in the preprocessing of the conversations. The average is far from the median for the number of *replies* and *users*, indicating some outliers (i.e., some popular conversations often initiated by influencers). Also, we noticed that the average of *other* languages is higher than Spanish or English due to outliers in the number of replies to a specific type of conversations (games or sports). Although tweets in other languages belong mostly to conversations in English, the initial tweet contains multimedia (images or video) or limited text that difficult language detection by the Twitter API. Figure 1 depicts the distribution of the number of replies on the conversations.

For further analysis, we perform several steps of pre-processing to the conversations. We filter out conversations non-alternating users' tweets (single tweet per user), as well as conversations with less than 3 tweets and more than 10 tweets[4]. We used 3 as the lower bound for the English dataset due to the limited number of conversations in that language. After the preprocessing, our dataset contains 518 Spanish and 172 English conversations.

4 ANNOTATION

We select the 518 Spanish and 172 English conversations including the replies for the annotation task, which account in total 2193 and 730 tweets Spanish and English respectively. Additionally, we randomly select tweets regardless if it is in a conversation until complete approximately 4000 tweets for each language. We rely on the multi-class taxonomy proposed by Imran et al. [13] for the annotation task. In the multi-class taxonomy, each tweet can belong to one of the following categories: a) Injured or dead people b) Missing, trapped or found people c) Displaced people and evacuations d) Infrastructure and utility damage e) Donation needs or offers or volunteering services f) Caution and advice g) Sympathy and emotional support h) Other useful information, and i) Not related or irrelevant

[4]The lower bound was set to allow for at least two replies in the conversation, and we define the upper-bound after finding that 84% of the conversations had replies ≤ 10

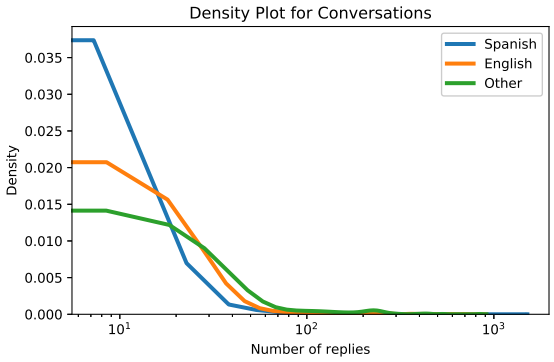


Figure 1: Distribution of the conversations by language.

We conduct our annotation process with two undergraduate students as human annotators and present them with the tweets selected for annotation. We provide to the annotators with instructions for the annotation task describing each of the categories and associated examples. For each tweet, we ask the annotators to select only one category based taxonomy. A third annotator resolves the discrepancies or disagreements between the two annotators by assigning the final category to the tweets without agreement.

Similar to the process in Imran et al. [13], we ask the annotator to make the judgments based on the text only, even if Twitter APIs truncated the text during the data collection. Also, we ask them not to open any link inside the text of the tweets as we seek to use only the available text for training the machine learning models. Since each tweet in our annotation process can have only one label, we calculate the percentage of agreement (p_o) between the two annotators. We also calculate the Fleiss' Kappa (κ) [9] agreement metric between the annotators. We calculate kappa metric as $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o define the empirical probability of agreement (i.e. the observed agreement percentage), and p_e accounts for a random agreement between annotators.

As shown in Table 3, we find that the per-label agreement varies from $\kappa = 0.75$ for lexically well-defined categories (such as *caution and advice* and *infrastructure damage*) and agreement $\kappa = 0.45$ for less clearly-defined categories (e.g., *displaced and evacuations).* The categories with few instances often have low scores because an error in the annotation affects by a large margin the agreement score. We also observe that some tweets can belong to several categories. In these cases, the annotators might differ in their judgment, and this lowers the agreement score even in categories with a large number of instances.

5 MODEL

In this section, we evaluate different approaches for modeling crisis-related tweets on multi-lingual settings for our annotated corpus. First, we describe the learning representation for the tweets, then the neural architecture for classification used, and finally the experimental settings.

5.1 Learning Representation

For tweets' text representation, traditional approaches require manually engineered features like cue words or sparse vector representation such as TF-IDF used in previous works [11]. In this work, we rely on the use of dense representation such as word embeddings Word2Vec [17]. Mikolov et al. [17] proposed an unsupervised language model using two log-linear models for computing dense representations from large (unlabeled) corpus efficiently: **a)** *bag-of–words* model CBOW that predicts the current word based on the context words, and **b)** a *skip-gram* model that predicts surrounding words given the current word. The approach show that both models can learn high accuracy syntactic and semantic regularities and overcome the issues in sparse representation models.

Word2Vec [17] represents each word in the corpus like an atomic entity and generates a embeddings vector for each word. In this aspect, Word2Vec and Glove [21] are similar; both define *words*

Table 3: Agreement statistics for Spanish and English tweets.

	Spanish				English			
Category	**N**	*po*	*pe*	κ	**N**	*po*	*pe*	κ
Injured or dead people	624	0.69	0.12	0.65	1,165	0.64	0.12	0.59
Missing or found people	30	0.77	0.27	0.68	4	0.75	-	0.75
Displaced people and evacuations	12	0.50	0.08	0.45	7	0.57	-	0.57
Infrastructure damage	157	0.64	0.07	0.61	73	0.75	0.11	0.72
Needs or offers	234	0.67	0.11	0.63	207	0.64	0.12	0.59
Caution and advice	61	0.77	0.08	0.75	39	0.64	0.18	0.56
Emotional support	451	0.67	0.10	0.63	325	0.62	0.10	0.59
Other useful information	753	0.69	0.10	0.66	426	0.59	0.13	0.53
Not related or irrelevant	1,846	0.69	0.11	0.66	1,946	0.66	0.11	0.61

as the smallest unit to train. However, Word2Vec does not take advantage of the *global context*. Both CBOW and Skip-Grams are predictive models and only use local contexts during training. In contrast, GloVe leverage the same intuition but uses a neural method to decompose the co-occurrence matrix into more expressive and dense word vectors.

N-gram feature is a critical improvement in FastText [18] compared to Word2Vec, and it aims to solve the out-of-vocabulary (OOV) issue. FastText enables word embeddings to encode sub-word information and produce more accurate vectors than Word2Vec.

Most recently, deep contextualized embeddings such as ELMo [22] and Flair [2], generate embeddings for a word based on the context, thus generating slightly different embeddings for each word depending on the context of its occurrence. We leverage on recent advances of the learning representation methods using the pre-trained embeddings and also fine-tuning the embeddings to our dataset.

5.2 Neural architecture

We are interested in identifying if a given tweet is related to a crisis event. We frame the problem of detecting crisis-related tweets as a multi-class classification task. To that end, we rely on both Convolutional Neural Networks (CNN) and LSTM sequence models using word embeddings as learning representation. Due to the unstructured, short, and noisy nature of the Twitter data, CNN models have shown to perform well for short-text classification task [20].

We train the CNN and LSTM models by optimizing the binary cross entropy loss using the adaptive gradient-based learning algorithm [16]. We set the learning rate and parameters to the values as suggested by the authors. We set the number of epochs to 10 for the case of the random embeddings and 5 when using the stacked embeddings with Glove-Flair [2].

We use dropout [25] after embeddings and hidden units to avoid overfitting. We initialized the word vectors in L with random embeddings in the case of CNN. In the case of LSTM, we use a vanilla version with random embedding and a Stacked Glove-Flair pre-trained embedding with LSTM.

5.3 Experimental Settings

Data Preprocessing: We normalize all characters to their lowercased forms, truncate elongations to three characters, convert every digit to `D`, anonymize twitter usernames to `userID`, and all URLs to `HTTP`. We remove all punctuation marks except periods, semicolons, question and exclamation marks. We tokenize the tweets using the NLTK toolkit [5].

Label Grouping: Due to the imbalance of the labels, we group the labels into a binary classification task to identify whether a tweet is related to a crisis event or not. A given tweet is related to a crisis if it belongs to any of the categories but *Not related or irrelevant*. The Table 4 shows the grouping labels that we use for the experiments. The grouping into two classes achieves a reasonable balance for both Spanish and English languages.

Table 4: Dataset for the binary classification task.

Label	**Spanish (es)**	**English (en)**
Crisis related	2322	2249
Not related	1846	1946
Total	4168	4195

Splitting Strategy: The splitting strategy for the dataset is to use 80% for training and development (using 10-fold cross-validation) and hold out 20% of the data as a test set. We randomly split the dataset into train and test sets to ensure that the classes distribution remains reasonably balanced in each set.

Classification Tasks: We define two set of experiments as described next. First, we train and test a model on the same language: train and predict on the Spanish dataset, and similarly train and predict on the English dataset. Second, we train a model on a language and predict in another language, e.g., train on Spanish dataset and predict on English dataset. We do not set up a third scenario that involves translating the dataset to a single language as described in Khare et al. [15] due to the cost constraint of using a translation API.

Classification Models: We implement a classifier based on Linear Regression (LR) as a baseline model and models based on neural architectures such as CNN and LSTM. Additionally, we evaluate an

Table 5: Performance of the models in single language and multi-lingual classification.

model	LR			LSTM			CNN			LSTM Stacked		
language	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
es-es	85.70	85.48	85.52	85.78	85.12	85.18	86.30	85.71	**85.77**	81.49	81.17	79.73
en-en	93.30	93.21	93.21	92.89	92.95	92.92	94.30	94.04	**94.05**	91.06	91.35	91.04
es-en	79.05	78.45	78.47	76.23	74.68	74.59	79.74	79.71	79.65	85.90	85.90	**85.88**
en-es	57.73	45.10	29.97	56.32	44.62	28.39	52.27	44.67	29.31	79.57	77.50	**77.49**

LSTM model that uses Stacked Embeddings (Glove word embeddings + Flair deep contextual embeddings).

Evaluation Metrics: We use *Precision*, *Recall*, and F_1 macro metrics to evaluate the performance of classification models, and we report the results on the test set.

6 RESULTS

The Table 5 shows the results of the classification task. The first column shows the source language in which is trained the model and the target language that predicts. Each of the models have three columns associated that represent the metrics *precision* (P), *recall* (R), and F_1 score.

For the first experiment, the CNN model outperforms other models training and predicting for a single language. The performance of the baseline model LR is on par with CNN, and better than the LSTM model that means TF-IDF based models are strong baselines in noisy short-text classification under a single language and single event but often fail to generalize for new events due to the OOV issue. The LSTM and CNN are using randomly initialized embeddings and fine-tuned during training which hinders the performance compared to using pre-trained embeddings. In this experiment, the LSTM with multi-lingual embeddings did not perform well, mainly because it does not apply the fine-tuning to the embedding weights.

The second experiment aims to train a single model for predicting tweets in another language (row 3 and 4 in Table 5). Traditional approaches as LR using TF-IDF fail to generalize, and the performance fall in the case *es-en* and drop drastically in the case *en-es*. The reason that the first case is not as critical as the second is not apparent and requires further analysis. There is a small percentage (4%) in the Spanish dataset with a different language, but not enough to affect the training of the model. In the case of the English dataset, there is 10% of tweets with a language different from English.

However, the model LSTM with multi-lingual Stacked Embeddings generalize well in this setting and outperform other models, which is promising for tackling the multi-lingual scenarios in detecting crisis-related tweets.

7 DISCUSSION

This section details some of the findings related to the crisis-related conversations, cross-lingual analysis, and current limitations in this work.

7.1 Crisis-related conversations

We explore the conversational nature of the interactions on Twitter and how we can improve extracting insights during crisis events by analyzing entire conversations. In this analysis, we focus on both Spanish and English conversations. By considering only the *root* or *initial* tweets in the conversation, approximately 21% of them are not related to the earthquake, while the rest of the root tweets belong to the other categories.

We look at the replies of root tweets not related to the crisis, and we found most of the replies (97%) are also not related to the earthquake, but there is a remaining 3% of replies annotated as donations, emotional support, and other useful information even when the root or parent tweet of the conversation is not related to the crisis. This small percentage indicates that we might lose some information even when the root tweet is not related to the crisis; thus we need to understand it in the context of the conversation.

A similar situation occurs in the replies to root tweets that are related to the crisis. In this case, there is a 24% of the replies that are annotated as not related to the crisis (in the case of English conversations up to 35%), when those tweets are part of conversations related to the crisis. In this case, these tweets could help to analyze the objective or outcome of the conversation semantically.

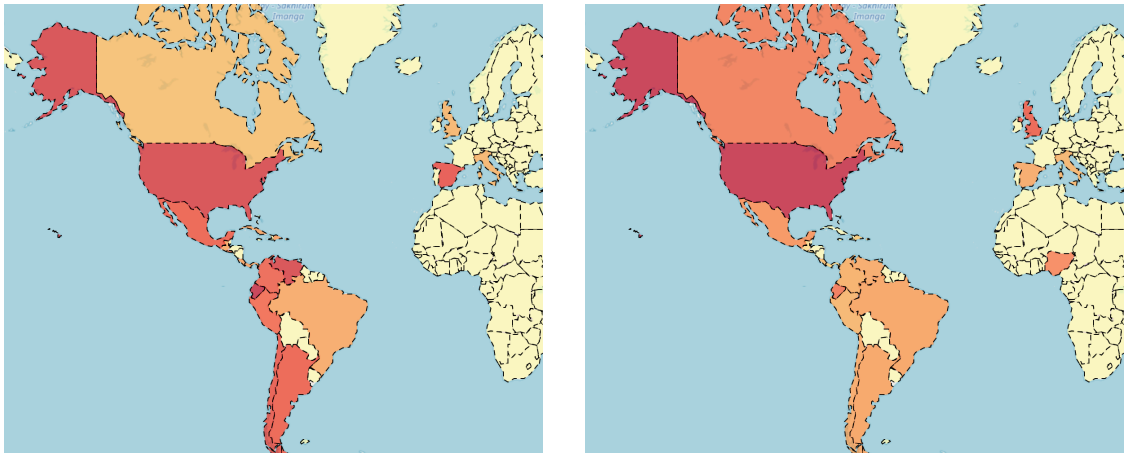
7.2 Cross-lingual analysis

Using the best model, we categorize the complete dataset to identify tweets related to the crisis event. We establish the country of the tweets' users by extracting that information from the *location* field in the metadata of the tweets. We preprocess the *location* field as it contains location information. Approximately 30% of the users do not provide location, and 15% contain noisy information in location. For our analysis, we consider only those with location information.

The Figure 2 shows the spatial distribution for each of the languages. In the Figure 2a, we show the crisis-related tweets in the Spanish language. For the Spanish language, the predominant country is where the earthquake occurred, Ecuador, followed by the United States, Venezuela. Also, there is a significant percentage in Spain, where there is a large number of Ecuadorian immigrants.

The Figure 2b shows that the Spanish-speaking countries decrease their participation in English dataset. The United States is the country where most tweets in English are posted, followed by England and Canada. An interesting issue would be to determine what percentage of the users are Ecuadorian immigrants living abroad.

Another aspect is the percentage of the annotated tweets associated to each category depending on the language. For the Spanish dataset, we identify the location information for approximately 49%



(a) Spanish crisis-related tweets **(b) English crisis-related tweets**

Figure 2: Spatial distribution of crisis-related tweets, darker color indicates larger percentage of tweets.

of the tweets. Most of the tweets (3.57%) related to *Needs / Offers* for goods and services came from the affected country, while 3.24% came from other countries. The majority of the tweets related to statistics about deaths and injured people come from other countries (13.87), while the tweets in the affected country are 2%.

For the English dataset, we identify the location information for approximately 77% of the tweets. The tweets related to *Needs / Offers* for goods and services that came from the affected country is almost inexistent, while 4% came from other countries. The majority of the tweets related to statistics about deaths and injured people come from other countries (29.46%), while tweets in the affected country are less than 1%.

7.3 Limitations

We identify several issues while conducting experiments that are important to mention. Due to the limited size of our annotated dataset, we do not claim to make conclusive observations in the analysis of crisis-related conversations, local vs. foreigners, and cross-lingual differences.

Given the conversational structure in the Twitter platform, the conversations might have several overlapping tweets, i.e., conversations can have several branches, and we treat each branch as an independent conversation, which could cause duplicated tweets in the dataset. In the dataset splitting strategy, we perform a random splitting, but further experiments should consider time-based splitting, as it is more similar to a real scenario. In this sense, models should be able to learn from incomplete conversations until more replies arrive. Using the context for learning to predict a new tweet could improve the performance of the classification models, we consider a future research direction, especially using attention mechanisms [28].

Due to the ambiguity in some tweets, assigning labels can be confusing even for human annotators, and a limitation in our annotation process is the number of annotation per tweet compared to previous works. Similar to previous works on crisis-related datasets, our corpus has a long tail of classes with few instances. Due to the imbalance in the labels, it is difficult for classification models to learn from very few instances. Some strategies to overcome the imbalance include the implementation of *Zero-shot* [30] or *Few-shot* [23] learning for labels with few instances.

The scarcity of annotated data in the Spanish language limits the capacity to use additional annotated data from other events to augment our dataset. Further evaluation of additional cross-lingual embeddings and language models are essential for transfer learning approaches to overcome the scarcity of annotated data in low-resource languages [10].

8 CONCLUSION

In this paper, we introduce an annotated corpus in Spanish and English language for the earthquake that occurred in Ecuador on April 16, 2016. We annotated not only isolated tweets but those that belong to conversations regarding the earthquake, which will enable future research for deeper understanding of that type of interactions and their outcome. We find that tweets often overlap semantically, an indication that a multi-label annotation would be more suitable for a more in-depth understanding of more complex interactions such as conversations. For our annotated corpus, we explore how we can identify crisis-related tweets on multi-lingual settings leveraging advances on multi-lingual deep contextual embeddings. We found that multi-lingual embedding outperforms other approaches based on sparse representation; however, for single language more simple model still perform better.

Future research directions include a comparative analysis of cross-lingual modeling of crisis-events using additional datasets publicly available. We also aim to improve the learning of labels with few instances, as in our corpus, through zero-shot or few show learning methods. Finally, we also seek to develop a fine-grained taxonomy as we detect approximately 15% of the tweets can belong to multiple labels.

REFERENCES

[1] Adam Acar and Yuya Muraki. 2011. Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities* 7, 3 (2011), 392–402.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*. 1638–1649.

[3] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain Adaptation with Adversarial Training and Graph Embeddings. *arXiv preprint arXiv:1805.05151* (2018).

[4] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Graph Based Semi-Supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets. In *Twelfth International AAAI Conference on Web and Social Media*.

[5] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[6] Gregoire Burel and Harith Alani. 2018. Crisis Event Extraction Service (CREES)-Automatic Detection and Classification of Crisis-related Content on Social Media. (2018).

[7] Mark A Cameron, Robert Power, Bella Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 695–698.

[8] Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

[9] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[10] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 328–339.

[11] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 67.

[12] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 159–162.

[13] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894* (2016).

[14] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. 2016. Cross-language domain adaptation for classifying crisis-related short messages. In *13th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2016*. Information Systems for Crisis Response and Management, ISCRAM.

[15] Prashant Khare, Grégoire Burel, Diana Maynard, and Harith Alani. 2018. Cross-Lingual Classification of Crisis Data. In *International Semantic Web Conference*. Springer, 617–633.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[19] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*.

[20] Dat Tien Nguyen, Kamla Al-Mannai, Shafiq R Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks.. In *ICWSM*. 632–635.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[22] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1. 2227–2237.

[23] Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3132–3142.

[24] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[25] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[26] Johnny Torres, Carmen Vaca, and Cristina L Abad. 2017. What Ignites a Reply?: Characterizing Conversations in Microblogs. In *Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. ACM, 149–156.

[27] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1619–1629.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[29] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. 2011. Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency.. In *ICWSM*. Barcelona, 385–392.

[30] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1004–1013.

Detecting Areas of Potential High Prevalence of Chagas in Argentina

Antonio Vazquez Brust
Fundación Bunge y Born

Tomás Olego
Fundación Bunge y Born

Germán Rosati
Fundación Bunge y Born, UNSAM

Carolina Lang
Fundación Bunge y Born, UBA

Guillermo Bozzoli
Fundación Bunge y Born

Diego Weinberg
Fundación Mundo Sano

Roberto Chuit
Fundación Mundo Sano

Martin A. Minnoni
Grandata Labs

Carlos Sarraute
Grandata Labs

ABSTRACT

A map of potential prevalence of Chagas disease (ChD) with high spatial disaggregation is presented. It aims to detect areas outside the Gran Chaco ecoregion (hyperendemic for the ChD), characterized by high affinity with ChD and high health vulnerability.

To quantify potential prevalence, we developed several indicators: an Affinity Index which quantifies the degree of linkage between endemic areas of ChD and the rest of the country. We also studied favorable habitability conditions for *Triatoma infestans*, looking for areas where the predominant materials of floors, roofs and internal ceilings favor the presence of the disease vector.

We studied determinants of a more general nature that can be encompassed under the concept of Health Vulnerability Index. These determinants are associated with access to health providers and the socio-economic level of different segments of the population.

Finally we constructed a Chagas Potential Prevalence Index (ChPPI) which combines the affinity index, the health vulnerability index, and the population density. We show and discuss the maps obtained. These maps are intended to assist public health specialists, decision makers of public health policies and public officials in the development of cost-effective strategies to improve access to diagnosis and treatment of ChD.

CCS CONCEPTS

• **Applied computing** → *Health care information systems*; • **Information systems** → *Data mining*.

KEYWORDS

Chagas disease; neglected tropical diseases; epidemics; health vulnerability; migrations; call detail records; social network analysis

ACM Reference Format:
Antonio Vazquez Brust, Tomás Olego, Germán Rosati, Carolina Lang, Guillermo Bozzoli, Diego Weinberg, Roberto Chuit, Martin A. Minnoni, and Carlos Sarraute. 2019. Detecting Areas of Potential High Prevalence of Chagas in Argentina. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3316485

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316485

1 INTRODUCTION

This document presents the processing criteria and the analysis techniques used to develop a map with high spatial resolution, which allows us to identify areas of Argentina that are more likely to be inhabited by a population potentially affected by Chagas disease (ChD). In the absence of disaggregated information regarding the prevalence of the disease in each locality, and to inform public policies aimed at treating it, this instrument will allow a focused follow-up in the geographic areas that need it most.

The fundamental dimension of analysis is the *potential prevalence of Chagas*, which will be operationalized through a *Chagas Potential Prevalence Index* (ChPPI). It is defined as the *proxy* indicator to determine whether a population lives in an area characterized by a high probability of being affected by *Trypanosoma cruzi*.

The index combines a set of characteristics of the population, such as their mobile phone communications, and a series of variables and indicators linked to health coverage in the areas of residence. As will be explained later, there are grounds to consider an index of this kind as a promising alternative in countries that do not have complete epidemiological records.

The first source of data used were anonymized Call Detail Records (CDR), which contain information about incoming and outgoing calls of users. This registry allowed the construction of an *Affinity Index*, which determines the level of linkage of the resident population in the Argentine territory, particularly in endemic areas (that is, where a high prevalence of the ChD by proximity and contact with the natural vector, *Triatoma infestans*, is observed) with the population living outside these areas.

On the other hand, the proximity of health centers and the socioeconomic level of the population as health determinants were used for the construction of a Health Vulnerability Index, which complemented the analysis of the Affinity Index. These determinants are closely related to the health status of a person, understood in a broad sense: biological, psychological and social health. If they are absent or of insufficient magnitude, a state of vulnerability is produced [7].

By combining both indices, we generated an indicator that seeks to identify areas characterized by high contact with areas where Chagas disease is endemic and where a high level of health vulnerability is registered. Complementing the affinity information with

health vulnerability information allows us to interpret the relationship between both and to understand holistically the phenomenon of infection and the local socioeconomic situation.

The information presented has a high level of spatial disaggregation. The minimum unit of analysis is the census block, the smallest statistical unit for which public socio-demographic information is available. The size of census blocks in urban areas is determined by the number of homes: the blocks cover an average of 300 homes. In Argentina, according to data from the 2010 Census, there are more than 52,000 census blocks distributed throughout the country.

The map, by pointing out the "hot" areas of high affinity with the endemic area and high health vulnerability, was conceptualized and developed as an input for different users: researchers, public health specialists, decision makers of public health policies and public officials, among other actors. In particular, for decision-makers, this map would facilitate the development of cost-effective strategies to improve access to diagnosis and treatment of ChD.

The document is structured as follows: Section 2 summarizes some relevant background on the subject and marks the main advances regarding the works of [9, 10, 28]. In Section 3 the methodology used for the construction of the affinity index is presented, together with the processing of the information linked to the CDRs. Section 4 presents some fundamentals about the concept of health vulnerability and a detailed explanation of the different procedures and techniques applied for the construction of the vulnerability index. In Section 5, the functional form and calculation of the *Ch-PPI* is developed and justified, and in Section 6 the final map is presented and a first descriptive analysis is made from the results. Finally, the limitations and future lines of research opened by this work are discussed in Section 7.

2 RELATED WORK

The continued emergence of large databases, added to the growing capacity of computer processing (the phenomenon called *Big Data*), has produced great advances in different disciplines, including epidemiology. By leveraging billions of cellular call records, and under certain conditions, it is now possible to model the diffusion pattern of certain endemics [12].

Call record databases have emerged as a source of particular interest for the study of human migrations [5]. A review of recent scientific literature shows an increasing body of work using call record analysis for the study of mobility patterns on different scales. The use of these new sources in mobility studies offers some notable advantages over traditional options such as origin-destination surveys.

Call records offer optimal geographical and population coverage [31], and with a much lower cost of acquisition than the significant resources required to carry out large-scale mobility surveys. As shown in [24], the analysis of millions of call detail records allows them to infer patterns of mobility aggregated at different geographic scales, disaggregating certain characteristics by demographic group. On the other hand, the still novel nature of this source implies that its use as a research resource is in its initial and exploratory period. Given that the data is being collected for purposes other than scientific research, the presence of undocumented biases, noise and omissions is expected. In particular, biases depend on the cellular telephony coverage in the analyzed area and the market penetration of the different mobile phone companies. In addition (in contrast to traditional mobility surveys) call records are extremely detailed in their spatial and temporal attributes, but much more superficial in the sociodemographic ones [31].

For this work, we had access to anonymized records of about 50 million calls per day between mobile phones, over several months. Distinct call patterns can be inferred, establishing their relationship with social phenomena such as seasonal or long-term migrations. Furthermore, if statistics were available at a detailed level (that is, at the highest possible level of disaggregation), the migratory patterns of individuals infected with parasites, viruses or bacteria could be inferred under certain circumstances.

In recent years, seminal analyses of mobile phone communications to detect potential risk areas for the Chagas disease were carried out in two Latin American countries (Argentina and Mexico). These works were presented in [9, 10, 28]. These studies showed that geolocated call records are rich in social information and can be used to infer whether an individual has lived in an endemic area at some point in his life.

In this study, progress was made in two fundamental directions with respect to the aforementioned works:

(i) The affinity model between endemic and non-endemic areas previously used was further developed. Instead of assuming uniform probabilities of Chagas transmission in the endemic area (EA), complementary information was used about the homes located in EA in order to identify differentials in those probabilities of transmission. This point is addressed in Section 3.

(ii) A more general dimension was incorporated which affects the probability of contagion of the disease: health vulnerability, developed in Section 4.

In the present work, the correlations found between communications, mobility, access to the health system, demographic characteristics and the distribution of Chagas disease are refined; and the project is scaled up at the national level.

3 AFFINITY INDEX

3.1 Motivation

As previously mentioned, the Affinity Index quantifies the degree of linkage between endemic areas of Chagas disease in Argentina (also known as the emphGran Chaco Argentino) and the rest of the country. This index integrates two differentiated dimensions that use different sources of data: telephone records, housing conditions, and location of health centers, among others.

The first dimension is linked to the affinity between endemic and non-endemic areas, that is, to what extent the different areas of the country are linked to endemic areas (EA). Given that outside the EA, there are no favorable conditions for the reproduction of the main vector of the disease (the species *Triatoma infestans* or "vinchuca"), one of the main vehicles for the spread of the disease is linked to migratory currents. This fact is even more relevant considering that many of the provinces that make up the Gran Chaco have been historical producers of labor for the Argentine economy, at least between the beginning and ends of the 20th century [4]. Therefore, the possibility of detecting areas with a high potential prevalence

of ChD requires the correct detection and mapping of migratory flows between the Gran Chaco and the rest of the country.

A first obstacle in this sense lies in the lack of demographic information with a level of disaggregation adequate for the objectives set. In effect, the classical source (and virtually the only source of the National Statistical System with national rural and urban coverage) for the study of internal migrations is the National Population Census. However, in general, they are limited to quantifying internal migration flows at the provincial level. Although the places of birth, habitual residence and previous residence (five years ago) are investigated, the data are published only at the provincial level.

These limitations entail the need to explore other sources of information to quantify migration flows. The quantification of the affinity between both types of zones is based on the assumption that a high connection of mobile communications (properly filtered and processed, as will be seen below) constitutes a valid approximate indicator of the existence of migration flows between both areas. A high degree of affinity between an endemic area and an external area could be considered as an indirect indicator of the presence of a migrant population from an endemic area that lives at the time of analysis in a non-endemic area.

In this way, quantifying the degree of affinity of each region with high spatial precision would allow identifying those areas where a population with a greater chance of having contracted Chagas resides, due to the dynamics of transmission in Argentina [6]. Thus, the infection would have occurred in a previous period, when said population lived in an endemic area, or transplacentally, as those areas with a population that comes from an endemic area will also be more likely to have infected women of childbearing age, who may be infected and transmit the disease to their progeny.

The second dimension constitutes an extension with respect to previous works [9, 10, 28], wherein the endemic area is considered as a homogeneous area. However, the Gran Chaco is not an area of homogenous vector transmission because of its geographical condition, history of control actions and socio-demographic and living conditions of its inhabitants. Although there are no complete and reliable records, the transmission (or the annual incidence) is facilitated in part by poor infrastructure of homes and the surrounding areas [8]. For this reason, we sought to identify the affinity between the human habitat and the vector of the disease (vinchuca).

3.2 Processing Mobile Phone Data

The main input used for the construction of the Affinity Index is mobile phone activity information. The database used consists of geolocated antennas and a table of calls (*Call Detail Records* or CDR), wherein each call is associated with the antenna used to make the connection.

The analyzed database has approximately 70,000,000 records per day, collected over a period of 5 consecutive months. To account for interactions between users, these were identified using an encrypted number. In this way, the anonymity of users in the database is protected, while maintaining the ability to distinguish between different users.

The CDRs are organized in registers, and each one provides the following fields: encrypted number of the originator, encrypted number of the destinatary, direction of the call (incoming or outgoing), date and time of the call, duration of the call in seconds, code of the cell tower used.

The CDR analysis and extraction of the affinity score with endemic Chagas areas, for each antenna, can be summarized in three sequential stages.

Detection of *home antenna* of each user. The first step consisted in the detection, for each user u, of its *home antenna* H_u, with the aim of geolocating the information of the communications. Thus, H_u is defined as the antenna in which most calls of the corresponding user are recorded in the period studied, based on the set of calls made on weekdays at night. That is, from Monday to Thursday, between 8pm at night and 6am in the morning of the following day. This range was chosen assuming that it coincides with the time slot during which most of the people remain in their residence. If there is more than one antenna with a maximum number of calls, one of them was randomly selected. The number of users for whom H_u was defined is around 15 million.

Calculation of seed affinity for each antenna. In this step, for each antenna a, the *seed affinity* s_a was determined. The affinity depends on the characteristics of the houses in the area where the antenna is located (see Section 3.3). Therefore, it depends only on geographic and demographic attributes, and not those related to phone communication.

As input to the process, a partition was made by quartiles of an indicator of the housing material conditions, whose estimation is developed in Section 3.3. Each antenna contained within the ecoregion of the Gran Chaco corresponds to an affinity indicator equal to the quartile of housing conditions to which it belongs (an integer between 1 and 4 inclusive), while outside the Gran Chaco polygon the indicator is 0. In this way, the s_a of each antenna can take an integer value between 0 and 4.

Calculation of affinity indicators for antennas using CDR. For each inhabitant u of any antenna in the country, the goal was to assign an affinity level based on their telephone communications in the social graph G. In particular, the set $V_u(G)$ of users that make up the *neighborhood* of u in the social graph was calculated. Each neighboring node $v \in V_u(G)$ modifies the u score according to the intensity of the edge (u, v), and the seed demographic indicators in the region of the antenna where v lives (i.e. s_{H_v}).

Given the modified indicators for each user, we can add the results grouping by household antenna, to find a distribution of the affinity indicator for each of them.

First, for each user u his seed affinity was defined as $s_u = s_{H_u}$ that is, each user was associated with the seed affinity of the area in which they live.

Then, given u and its neighborhood $V_u(G)$, the affinity indicator s'_u of each user was defined as $s'_u = \max_{v \in V_u(G)}\{s_v\}$. This means that the affinity indicator modified by the social graph is obtained by calculating the seed affinity for each $v \in V_u(G)$, and then taking the maximum of the affinities that were found. Note that this number is also an integer between 0 and 4.

In this way, a series of indicators was defined for each antenna a, which is the information of how many users among the inhabitants of a have each of the values s'_u. We call $H_{a,k}$ the subset of the

inhabitants H_a of an antenna a, such that its s' is k, that is:

$$H_{a,k} = \cup_{(h \in H_a / s'_h = k)} \{h\}.$$

By reporting the number of users in $H_{a,k}$ for each of the values $k \in 0, \ldots, 4$, we obtain a summary of the distribution of affinity levels in each antenna a in particular. That is, for each antenna, the distribution of users is given by the tuple:

$$< a, |H_{a,0}|, |H_{a,1}|, |H_{a,2}|, |H_{a,3}|, |H_{a,4}| >$$

3.3 Housing Conditions

Given that it is possible to relax the assumption about the uniform transmission chances throughout the ecoregion of the Gran Chaco, and considering the social component of the disease, the habitability conditions of the houses and their direct relationship with the conditions conducive to housing the vector, we propose to disaggregate the region into multiple sub-areas. Each of the subdivisions represents a quartile with respect to an indicator that measures the degree to which local housing materials favor the presence of *Triatoma infestans*.

For this, we used information from the 2010 National Population and Housing Census [20] and we constructed an index that quantifies for each house its viability conditions for lodging the *vinchuca*.

The ranch houses whose construction materials include adobe walls without plaster and roofs of cane, mud or straw favor the domiciliation of the vector [19]. We worked with the following variables:

- Predominant material of the floors
  - Ceramic, tile, mosaic, marble, wood, carpeting
  - Cement or fixed brick
  - Soil or loose brick
  - Other
- Predominant material of the roof exterior
  - Asphalt cover or membrane
  - Tile or slab (without cover)
  - Slate or tile
  - Metal sheet (without cover)
  - Fiber cement or plastic sheet
  - Cardboard sheet
  - Reed, palm, board or straw with or without mud
  - Other
- Internal ceiling (yes / no)

The favorable habitability conditions for the *vinchuca* are the following: (i) Predominant material of floors is "Soil or loose brick"; (ii) Predominant material of roofs is "Reed, palm, board or straw with or without mud"; (iii) Absence of internal ceiling.

Since these are discrete and categorical measurement level variables, we used the Multiple Correspondence Analysis as a dimensionality reduction technique to recover the latent variables; then the dimension with the greatest variability was obtained and the coordinates for that latent variable were recovered.

3.4 Antenna Level Indices

The information of the polygon of the endemic area has a level of resolution associated to the polygon of the census radius. Now, the information to generate the contact graph is associated with the antennas, that is, fixed points in space. Therefore, the first step for the generation of the antenna level indicators was to estimate the coverage area of the antennas in Argentina. For this purpose, a Voronoi diagram was used (shown in Fig. 1).

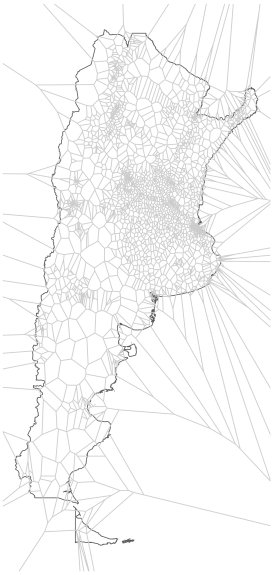


Figure 1: Voronoi cells generated based on the position of the analyzed antennas

The results of the housing quality index were added by each Voronoi cell, assigning a number of households based on the percentage of the area that intersected each census block. In other words, each block contributed a percentage of its values based on the proportion of surface shared with the Voronoi polygon or intersecting polygons. Finally, the values obtained by antenna were graduated according to quantiles.

4 HEALTH VULNERABILITY INDEX

It is possible to state that the chances of contracting ChD, or of being infected by *Trypanosoma cruzi*, are associated not only with the probabilities of contact with the vector or with movements or contacts with endemic areas. There are also determinants of a more general nature that could be encompassed under the concept of "Health Vulnerability." These determinants are associated with access to services and health coverage of different segments of the population. By developing an index that quantifies this dimension, we can complement the Affinity Index with endemic areas of Chagas, since estimating the Health Vulnerability allows to prioritize "hot zones" outside the endemic area, those that have low access to sanitary services.

4.1 Definition

The existence of disparities in access to health services is a phenomenon that has been studied and documented on multiple occasions, highlighting evidence of disparities in access and constant medical coverage for different strata of the population. In general, the most

impoverished population segments and/or residents in isolated areas have lower levels of access to these health benefits [3, 30].

In the literature on health problems, this notion of "unequal access" is often differentiated from the so-called "vulnerability" that refers, ultimately, to the risk of developing certain diseases or the fact of being exposed to certain environmental risk factors. In this sense, the study of *vulnerable populations* or *vulnerability factors* is of interest.

An obstacle to the quantification of "vulnerability" emerges from the need to consider multiple factors that could explain inequalities in access to the health system. The main indicators used by the studies analyzed in [15] are the condition of poverty, belonging to ethnic or racial minorities, the presence of chronic mental or physical illnesses and the lack of medical care. It should be noted that these indicators are defined at the individual level. There are also other factors that determine the level of health vulnerability that are more linked to the environmental dimension [27].

In our approach, several of the aforementioned indicators were considered. The proposed notion of health vulnerability is composed of the following associated factors:

(i) *Access to health services and benefits from the state:* for this dimension, the proximity to health providers was used as the main indicator. A dataset was constructed that contains the location (latitude and longitude) of the vast majority of state health providers across the country. For this, sources from the National State and the Provincial States were integrated. The walking time was calculated from various points to the nearest Health Center.

(ii) *Socio-economic Index of the population (SEI):* to construct the SEI, we used census information. Although this point is detailed below, it can be mentioned that the calculation of the indicator involved the processing of census information, corresponding to the CNPyV of year 2010. To this end, a series of relevant variables (educational level, indicators of unsatisfied basic needs, etc.) were selected and combined using *variational autoencoders*, a method for dimensionality reduction based on neural networks (see Section 4.4).

4.2 Methodology of Construction

The dimensions mentioned above were combined to build a *Health Vulnerability Map*. The objective of the vulnerability map is to identify areas with a potential deficit in the health coverage of the population, that is, that do not meet a minimum threshold in access to health services. Taking into account this objective, a metric was constructed that allows ordering and classifying the different zones according to this potential deficit.

For the construction of the map, the following sources were collected and analyzed: census data [20]; census block polygons; location of public health providers: public hospitals, health centers and sanitary posts (in total, 16,564); and the axes of streets (national and provincial routes, roads and urban traces) used to calculate by simulation distances between homes and health providers.

As mentioned, although data were processed at a lower disaggregation level (such as individual census data or data from health providers), for the construction of the final map, this information was added at the census block level .

The following sections detail the different procedures used for the preparation of the information to exploit population strata and the different processing and analysis techniques used.

4.3 Accessibility to Hospitals and Health Centers

Construction and cleaning of the Health Providers dataset. For the construction of the indicator *Closeness to Health Providers*, the first step was the construction of a dataset with location records of the greatest possible number of health providers in the whole country, located with latitude and longitude coordinates. This dataset was built from the integration of different sources of official data:

(i) National Base of Hospitals and Primary Care Centers: it was compiled by the Argentine Health Information System (https://sisa.msal.gov.ar/), and published by SEDRONAR on the IDERA site (http://catalogo.idera.gob.ar). This dataset was used as the starting point and *master base*. It was enriched and corrected based on information obtained from additional sources.

(ii) SUMAR program health providers. The information published on the site http://programasumar.com.ar/efectores/ was collected via scraping.

(iii) List of hospitals and health care centers of the National Program for Sexual Health and Responsible Procreation (Ministry of Health). Data available by province. The data was downloaded and georeferenced. Source: http://www.msal.gob.ar/saludsexual/centros.php.

Classification of health providers by level of complexity. Subsequently, and as the last stage of the dataset cleaning process, the health providers were classified according to their level of complexity. We sought to reflect the fact that proximity to a highly complex hospital implies access to health benefits a priori greater than the proximity to a small health post. The type of problems, the emergencies attended and the attention provided by these establishments differ markedly. There were different classifications in the datasets used related to the notion of complexity, and the criteria varied among sources: the classifications were not homogeneous in the different lists of health providers consulted.

A revision work was carried out with experts from the Mundo Sano Foundation, which allowed us to arrive at a classification that unifies the different denominations, producing a simple classification into three categories, in decreasing order of complexity: (1) Hospital; (2) Health Center; (3) Sanitary Post.

After discarding the public health effectors that do not belong to any of the defined categories (for example, geriatric or administrative offices), 15,903 records of the 16,654 of the total collected were preserved.

Computing travel time to the nearest health center. The next step was to compute the time necessary to reach the closest health care provider. From this point, the information was aggregated at the census block level.

We wanted to find the nearest public effector for each census block. Since the shape, boundaries and surface of the census blocks are very dissimilar throughout the country (especially in rural or sparsely populated areas), we decided to calculate the distances and times in the following way:

1. Within each block, 5 points (pairs of coordinates) were selected at random.
2. The nearest health effector was identified for each point, using k-NN (k nearest neighbors).
3. The distance / time to the nearest health provider was calculated for each point.
4. The 5 distances / times were averaged to obtain the final value.

This procedure was performed for each category of health providers: Hospital, Health Centers and Sanitary Posts.

For traveling time computations, we used Open Source Routing Machine (OSRM), a high-performance routing system that indicates the shortest route through public roads between any pair of source-destination coordinates [18]. To determine the routes, OSRM uses street grids downloaded from Open Street Map [26], a public repository of geographic information whose data quality has established it as a frequent source for mobility studies [16, 23].

Walking times. The indicator used to measure access to health coverage was the walking time to the nearest health provider. This indicator is relevant given that there is evidence that, at least for certain types of medical treatments, walking distance to a health facility is a good predictor of the chances of completion of such treatment. Indeed, the study [3] suggests that a distance greater than one mile (approximately 1.6 km) results in a considerable increase in the probability of not completing a rehabilitation treatment. In turn, distances greater than 6.4 km result in a decrease in the average duration of treatment in almost two weeks.

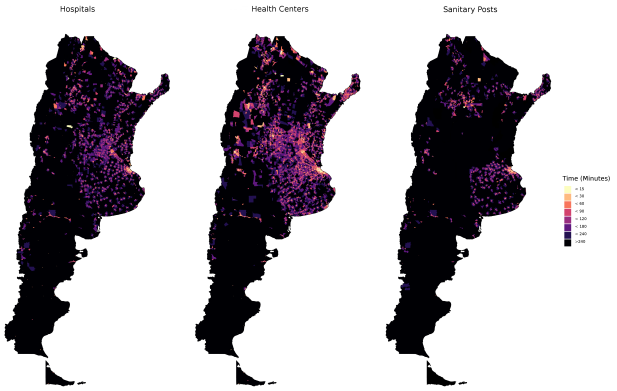


Figure 2: Walking times to health providers. Aggregated by census block.

Fig. 2 shows maps at the census block level of walking times to Hospitals, Health Centers and Sanitary Posts for the whole country. The final distance for each block Δ_r is the median distance between all the points sampled in each block, and includes all the distances to health providers in a census block r.

4.4 Socio-Economic Index

Input data. To compute the *Socio-Economic Index* (SEI), we used data from the 2010 Census. We worked with data at the individual level. Given that the SEI is usually a variable measured at the household level, we decided to calculate an index for each head of household in the Census dataset.

Table 1: Indicators used for the construction of SEI.

Variable	Unit
Condition of home ownership	Housing
Quality of Materials	Housing
Quality of Connection to Basic Services	Housing
Quality of Construction	Housing
Overcrowding	Household
Unsatisfied Basic Needs (UBN) indicator	Household
Educational level of the Household	Household
Number of Unemployed in Household	Household
Existence of domestic services	Household
Activity condition	Individual (head)
Educational level	Individual (head)

To estimate the values of the index, we used the variables from Table 1, which were ordinalized. To build the index, a thermometer encoding was used for the ordinal variables. Let N be the number of cases and $v_1, \ldots, v_I$ the variables. For each variable v_i, there are K_i categories. The following coded variables were created $x_{k_i}^{(i)}$ for each variable v_i and for each category k_i where $2 \leq k_i \leq K_i$. In each case j with $1 \leq j \leq N$ it holds that:

$$x_{k_i}^{(i)}(j) = \begin{cases} 0, & si\ v_i(j) < k_i \\ 1, & si\ v_i(j) \geq k_i \end{cases} \quad (1)$$

Construction of the final SEI. For the construction of the final index, we used a dimensionality reduction technique called *autoencoder* [14]. Autoencoders are an architecture based on neural networks. In general, an autoencoder has the objective of finding a representation of the input data (*encoding*), usually in order to perform a reduction of dimensionality. Autoencoders work by simply learning to replicate the inputs in the outputs. While this seems a trivial problem, by introducing various restrictions to the network, this task can become very complex. For example, you can limit the size of the internal representation or add noise to the inputs [13].

An autoencoder is composed of two parts: an encoder (or *recognition network*) that converts the inputs to an internal representation, followed by a decoder (or *generative network*) that reconverts the internal representation to the outputs.

It usually has an architecture similar to a Multi-Layer Perceptron, with the restriction that the number of neurons in the input and output layers must be equal. The trained model has as final layer a logistic function and performs a *dropout* of 0.5 in its intermediate layers to regularize and achieve coefficients with good generalization capacity (see Fig. 3).

The model was trained with ADAM [13] defining as data batches resampling with repetition on the empirical distribution of cases, to favor convergence.

As we worked with the whole population (since the data set is the census itself), the objective of the model was the generation of a descriptive measure. That is why the ability to explain the same population using the weighted average of the probability of each

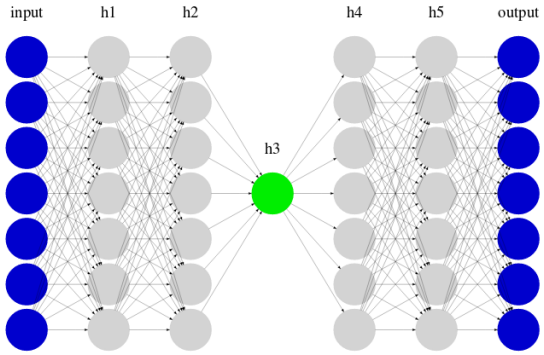


Figure 3: Scheme of the autoencoder used.

variable category was taken as a metric:

$$E(\hat{x}, x) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{k_i=2}^{K_i}\frac{1}{\sum_\ell K_\ell}e^{x_{k_i}^{(i)}\log\left(\hat{x}_{k_i}^{(i)}\right)+\left(1-x_{k_i}^{(i)}\right)\log\left(1-\hat{x}_{k_i}^{(i)}\right)}$$

The final model retains 85 % of the total input information. In this way, starting at h_3 each head of household, and therefore, each household, is classified with a value resulting from the autoencoder which we will call s_i, the SEI.

From the SEI, an aggregate measure was generated for each block, based on the socioeconomic level of the heads of household. Thus, for the heads of household i with a socioeconomic level index s_i that live in the census block r with a population of n_r heads of household, we define a variable η as

$$\eta_r = \frac{1}{4}Q_{.25}(\mathbf{s}_{r_{1,\dots,n_r}}) + \frac{1}{2}Q_{.5}(\mathbf{s}_{r_{1,\dots,n_r}}) + \frac{1}{4}Q_{.75}(\mathbf{s}_{r_{1,\dots,n_r}}) \quad (2)$$

where Q_p is the corresponding p quantile and η_r is the result of applying the summary measure known as "Tukey Trimean" which achieves a compromise between robustness and efficiency compared to the median.

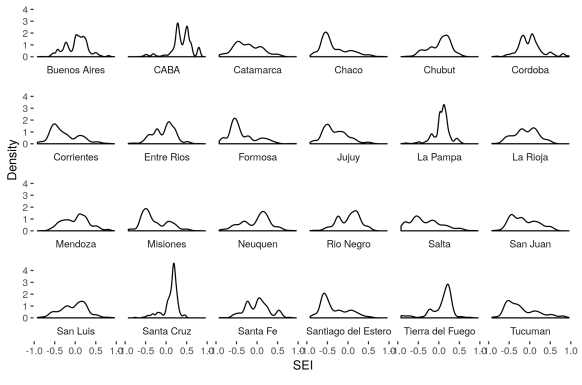


Figure 4: Density plot of the SEI by province.

As can be seen in Fig. 4, the constructed index seems to capture to a large extent the provincial disparities: the CABA presents a distribution clearly skewed to the right (higher values of the SEI); on the other hand, provinces such as Chaco, Formosa, Jujuy and Salta have distributions skewed to the left.

4.5 Generating the Health Vulnerability Index

To build the final Health Vulnerability Index, it was necessary to combine the socio-economic index and the accessibility to health centers at the census block level. The strategy adopted to generate a composite index was that of principal components, that is, to look for the linear combination of the variables that can explain the maximum variance.

Now, when the distribution of the variables presents atypical distributions, that is multimodal, asymmetric and/or with heavy tails, the interpretation of the main components is difficult because the method is sensitive to the scale of the variables [22]. A possible solution is to transform the variables using the data ranks [1, 2]

Therefore, following [29], we sought to standardize each of the variables X_j by means of the rankit transformation:

$$rankit(X_{i,j}) = \frac{r_j(X_{i,j}) - 0.5}{n} \quad (3)$$

For each observation i of each variable X_j the rank $r_j(X_{i,j})$ (which varies between 1 and n) is calculated, then we subtract 0.5 and divide by the total of records n. In this way, the differences in units that could exist between variables were eliminated, and invariance was achieved respect to changes of scale, displacements and monotonous transformations.

Then, following [11, 17], a semiparametric principal component analysis was performed (see Algorithm 1).

Algorithm 1 Semiparametric Principal Components Analysis

procedure S-PCA(X)
  for `j in variables` **do**
    for `i in cases` **do**
      $z_{i,j} = \Phi^{-1}\left(\frac{r_j(x_{i,j})-0.5}{n}\right)$ ▷ Φ^{-1} is the inverse of the gaussian f.d.a.
    end for
  end for
  compute Σ ▷ Covariance matrix
  find U, S such that $\Sigma = \frac{1}{n}US^2U^t$ ▷ SVD
  return ZU^t, S ▷ Coordinates and eigenvalues
end procedure

To perform the combination of both variables, the Spearman correlation was calculated on the rankits [17] and the nonparametric correlation matrix was decomposed. We found that the main eigenvector absorbed 72% of the variability of the rankits, and that it was oriented in the inverse direction of mutual growth. This first eigenvector was thus considered as the optimal combination between both variables. In this way, and through the successive application of nonparametric transformations, we generated an index that is pollution tolerant and, above all, independent of the measurement characteristics of each input variable.

As will be seen below, the interest lies in having an index between 0 and 1 whose distribution is homogeneous for all units of analysis. A transformation of the cumulative distribution function type estimated by logsplines [25] was again applied using AIC as

a regularization criterion on the main direction, thus forming the index HV_r for each block r.

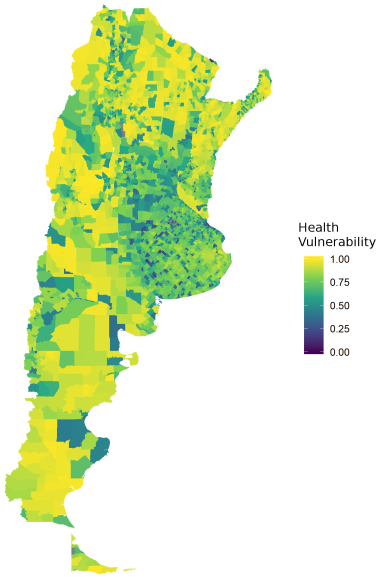


Figure 5: Health Vulnerability Index at the census fraction level for Argentina.

Figure 5 shows a visualization of the Health Vulnerability Index for the whole country, aggregated by census fraction. A census fraction is composed of a set of nearby census blocks [21].

We can detect two big zones: (i) a "central" region, essentially the Buenos Aires metropolitan area and the large urban agglomerations of each province, characterized by lower values of health vulnerability; (ii) the rest of the country, mainly the less densely populated areas, with higher values. However, even in the central region, there are zones with critical values.

5 GENERATING THE COMBINED INDEX

5.1 Population Density as Scaling Factor

The ultimate goal of the Chagas Potential Prevalence Map is to guide actions and interventions by the State (at different levels). This implies that the concentration of the population in each of the census blocks is a scaling factor that needs to be considered in the final index.

For the design of cost-effective intervention strategies, it is necessary to quantify the potential impact in terms of the number of people affected. In this sense, it may be preferable to carry out actions in areas with a slightly lower potential prevalence index, but with a greater concentration of population in the territory: the cost per person treated may be lower in these areas of high population concentration, minimizing transportation costs.

That is why population density represents the final component of the index. In effect, for each block r, density was calculated as the quotient between the number of inhabitants in the radius h_r and the total area in km^2 of the census block a_r. In order to combine it in the final index, the standardization strategy was again used by evaluating the cumulative probability distribution F on the variable, which removes both the unit differences and achieves invariance against changes in scale, displacements and monotonous transformations. The cumulative density function was estimated using logsplines:

$$d_r = \hat{F}_d\left(\frac{h_r}{a_r}\right) \tag{4}$$

5.2 Chagas Potential Prevalence Index

For the final construction of *ChPPI*, the following variables were taken into account: (i) local affinity with endemic area of Chagas, (ii) health vulnerability, and (iii) population density.

From the point of view of public policies, with limited funds, it makes sense to prioritize areas with high population density and/or high "affinity". This composite index would allow, properly calibrated, to establish an order in which to expand health policies attending ChD, that take into account these different factors.

The final index was composed as follows:

$$ChPPI_r = \frac{{HV_r}^{\alpha}\ {d_r}^{\beta}\ AI_r}{\frac{1}{R}\sum_{r=1}^{R} {HV_r}^{\alpha}\ {d_r}^{\beta}\ AI_r} \tag{5}$$

where ${HV_r}^{\alpha}$ moderates the effect of the Health Vulnerability component with α being the parameter that determines the impact; ${d_r}^{\beta}$ penalizes depopulated areas and β functions as the regulator; and AI_r is the Affinity Index for block r. In the denominator, R is the total number of census blocks.

Note that $0 \leq \alpha, \beta$. When $\alpha, \beta \longrightarrow 0$, the effect of the Health Vulnerability or population density (respectively) are canceled. On the other hand, when $\alpha, \beta \longrightarrow \infty$, the index is dominated by Health Vulnerability or population density (respectively).

Regarding $ChPPI_r$, it should be interpreted as a relative index since its values indicate how much higher the potential prevalence of Chagas is in a census block compared to the median population value.

6 MAPS AND RESULTS

The analysis of the results presented in this section intends to detect areas of interest for future field work and validation of the potential prevalence estimation model. That is, to try to find those areas in the country where the Chagas Potential Prevalence Index is high. This indicates a high affinity –activity of cellular calls– with endemic areas and high values in the health vulnerability index as an additional factor to prioritize a subsequent in-situ intervention. Both dimensions are combined in the final index.

For this analysis, only areas outside the Gran Chaco eco-region were considered –those where a high level of communication with the Gran Chaco is not attributable to the expected volume of calls among physically close persons.

However, we must take into account that the original map unit (census blocks) is not viable for the organization of field work: selecting relevant census blocks (with high affinity) and their corresponding contrasts would lead to a large territorial dispersion of the areas to visit. That is why the information was added at the local level.

Thus, after filtering blocks with less than 350 inhabitants, two metrics were built for each location: (1) the average affinity along all the blocks that make up the locality (weighted by the population in each block); (2) the average affinity of the blocks of the locality that were between the highest values of the national affinity distribution.

These metrics give rise to three types of possible situations: (1) localities that have high affinity but "dispersed", exhibiting a similar concentration in all the blocks that compose it; we will call them *high mean affinity*; (2) localities that have high and highly concentrated affinity values: that is, few blocks of the locality have high affinity values; named as locations of *extreme blocks*; (3) locations that meet both criteria.

Such are the fundamental guidelines followed for the analysis. In addition, within each locality, we discarded the census blocks whose the population density does not reach 350 inhabitants per km^2, since it is necessary to reach a minimum local population concentration to guarantee the operative viability of a possible field intervention.

Within each province we selected localities of the above mentioned types. **High affinity localities:** we selected the three with the highest average affinity weighted by the population of each block in each province. **Localities of extreme blocks and localities with both types:** the same criterion was followed, the difference is that in this case, the average is calculated on the blocks that survive the previous filter; this may cause that some provinces do not present localities of these types.

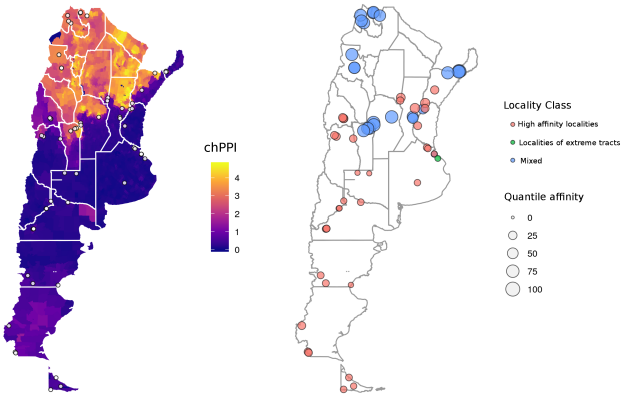


Figure 6: Localities (homogeneous and concentrated) selected as potential areas of intervention according to levels of Potential Prevalence Index (ChPPI) and according to affinity percentile.

Figure 6 shows the selected localities, projected on the map of the Potential Prevalence Index. As established in [10], the areas with the highest affinity indexes are concentrated in the Metropolitan Region of Buenos Aires (including the city of La Plata) and in the provinces of Patagonia – Neuquén, Río Negro, Chubut, Santa Cruz and Tierra del Fuego. Some localities of border areas to the Gran Chaco also appear as relevant, in provinces such as Misiones, Entre Ríos and Mendoza.

A first point to note is that there is only one location (La Plata, Buenos Aires) that is characterized by presenting extreme blocks but not high average affinity. At the same time, it can be noted that between the border areas of the Gran Chaco, those that combine extreme blocks and localities of high average affinity predominate within the selected localities. These localities are characterized (as expected) by higher values in affinity percentiles.

On the other hand, in the central and Patagonia zones, the situation seems to be different: the selected localities are characterized by being of high mean affinity and by lower values in affinity percentiles.

Although they share high affinity levels, the zones do not exhibit similar patterns in terms of their health vulnerability. In this dimension, the variability is high, covering areas in the entire range from low vulnerability in relation to the rest of the country (CABA, Vicente Lopez, La Plata), up to high levels: El Durazno and Jacipunco (Catamarca), Pueblo Libertador (Corrientes), Coranzulis (Jujuy), Poscaya (Salta).

Figure 7: From left to right, and from top to bottom: Candelaria (Misiones), Arroyito (Córdoba), Recreo (Santa Fe), 28 de Noviembre (Santa Cruz), Lima (Buenos Aires), Tolhuin (Tierra del Fuego).

Among the selected localities, in Fig. 7 we zoom in on 6 localities with a population greater than 5,000 habitants, each in a different province and in decreasing order of affinity. As expected, due to its geographical proximity to the Endemic Area (EA), the localities in Misiones, Córdoba and Santa Fe show a greater affinity compared to their counterparts in the Center and South of the country. However, we observe that the affinity does not decrease homogeneously, i.e. in a continuous gradient as the localities move away from the endemic area. On the contrary, localities were detected in the Province of Buenos Aires and in Patagonia whose degree of affinity is much higher than population centers in provinces closer to the EA such as La Pampa. This suggests the existence of considerable migrations from endemic regions to the highlighted localities.

7 DISCUSSION AND FUTURE WORK

A first result of the generation of this map involves opening some lines of work that may be of interest. The most obvious of them involves a process of revision, improvement and update of the data sources.

On the one hand, the sources of information used could be expanded by incorporating records obtained over several years to

identify dynamics related to migration patterns. In this work we used the volume of calls from residents of non-endemic areas with the endemic area as a proxy indicator of the existence of movements and population contacts between both types of zones. This implies an inferential leap: it is assumed that communication is a good indicator of the existence of migratory flows from endemic areas. In turn, a second assumption is that part of the migrant population is or was infected, at least, with a probability comparable to that of the non-migrant population.

Although this is a reasonable assumption, it is also expected that the same CDRs will allow –based on new processing– to estimate effective migratory patterns and volumes. That is, from the succession of user connections to different antennas, real population movements could be mapped between endemic and non-endemic areas.

On the other hand, the CDR data have limitations that need to be pointed out. Perhaps the most important one is linked to the degree of coverage of the mobile phone antennas. In effect, this variable alters the spatial granularity of the analysis. In areas of low density, the greater distance between antennas makes the area assigned to each one more extensive, and increases the difficulty of identifying the underlying census block. Therefore, it is recommended as future work to perform a spatial analysis of conglomerates in order to identify homogeneous areas of blocks to mitigate said problem.

Regarding the use of additional sources integrated in the Map (especially the Health Vulnerability), a second iteration in the cleaning and consolidation of the georeferenced data of health providers emerges as a necessary task. This task would imply the incorporation of sources that were not used in this first approach. In turn, the classification according to complexity level of health care providers deserves a review.

The incorporation of new sources of data and dimensions linked to environmental risk and vulnerability are also of potential interest for future work.

Another valuable result of this work is to generate a replicable work methodology, that can be applied in other areas. The quantification of interactions between dispersed geographic areas, as well as the indicators of Health Vulnerability, can be considered as transversal dimensions that affect the evolution and the transmission of other pathologies besides the ChD. Therefore, a second line of future work is the possibility of considering maps such as the one described here (and subsequent updates) as an input for the study of other infectious diseases.

REFERENCES

[1] John Bacon-Shone. 1992. Ranking Methods for Compositional Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41, 3 (1992), 533–537. http://www.jstor.org/stable/2348087

[2] M. J. Baxter. 1995. Standardization and Transformation in Principal Component Analysis, with Applications to Archaeometry. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 44, 4 (1995), 513–527. http://www.jstor.org/stable/2986142

[3] K. Beardsley, E. Wish, D. Fitzelle, K. O'Grady, and A. Arria. 2003. Distance traveled to outpatient drug treatment and client retention. *Journal of Substance Abuse Treatment, 25* (2003).

[4] Gustavo Busso. 2007. Migración interna y desarrollo territorial en Argentina a inicios del Siglo XXI. Brechas e impactos sociodemográficos de la migración interna interprovincial. In *IX Jornadas Argentinas de Estudios de Población. AEPA*.

[5] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical* 41, 22 (2008), 224015.

[6] Roberto Chuit, Ricardo E Gurtler, Laura Mac Dougall, Elsa L Segura, and Burton Singer. 2001. Chagas Disease-Risk assessment by an environmental approach in northern Argentina. *Revista de Patologia Tropical* 30, 2 (2001), 193–208.

[7] Roberto Chuit, Laura Mac Dougall, Maria Cristina Evequoz, Alicia Maria Bressan, and Maria Frias. 2008. TRAVEL AND INFECTIOUS DISEASE. *Revista de Patologia Tropical* 32, 1 (2008).

[8] R. Chuit, E. Subias, Perez AC., I. Paulone, C. Wisnivesky-Colli, and E.L. Segura. 2003. Usefulness of serology for the evaluation of Trypanosome Cruzi transmission in endemic areas of Chagas's Disease. *Revista da Sociedade Brasileira de Medicina Tropical* 22, 3 (2003), 119–129.

[9] Juan de Monasterio, Alejo Salles, Carolina Lang, Diego Weinberg, Martin Minnoni, Matias Travizano, and Carlos Sarraute. 2016. Analyzing the spread of Chagas disease with mobile phone data. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. https://doi.org/10.1109/asonam.2016.7752298

[10] Juan de Monasterio, Alejo Salles, Carolina Lang, Diego Weinberg, Martín Minnoni, Matías Travizano, and Carlos Sarraute. 2016. Uncovering the Spread of an Infectious Disease with Mobile Phone Data. In *Simposio Argentino de GRANdes DAtos (AGRANDA 2016)*.

[11] Bernhard Egger, Dinu Kaufmann, Sandro Schönborn, Volker Roth, and Thomas Vetter. 2016. Copula eigenfaces. In *11th International Conference on Computer Graphics Theory and Applications (GRAPP)*, Vol. 1. 50–58.

[12] Enrique Frias-Martinez, Graham Williamson, and Vanessa Frias-Martinez. 2011. An agent-based model of epidemic spread using human mobility and social network information. In *Privacy, Security, Risk and Trust (PASSAT) and Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 57–64.

[13] A. Gerón. 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly.

[14] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[15] C. Grabovschi, C. Loignon, and M. Fortin. 2013. Mapping the concept of vulnerability related to health care disparities: a scoping review. *BMC Health Service Research, 13, 94* (2013).

[16] Mordechai Haklay. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design* 37, 4 (2010), 682–703.

[17] Fang Han and Han Liu. 2014. Scale-Invariant Sparse PCA on High-Dimensional Meta-Elliptical Data. *J. Amer. Statist. Assoc.* 109, 505 (2014), 275–287. https://doi.org/10.1080/01621459.2013.844699

[18] Stephan Huber, Christoph Rust, et al. 2016. Calculate travel time and distance with OpenStreetMap data using the Open Source Routing Machine (OSRM). *Stata J* 16 (2016), 416–423.

[19] INDEC. 2003. Habitat y Vivienda por medio de datos censales. Calidad de los materiales de la vivienda (CALMAT).

[20] INDEC. 2011. Censo Nacional de Población, Hogares y Viviendas 2010. Base de datos. Definiciones de la base de datos.

[21] INDEC. 2011. Unidades Geoestadísticas. Cartografía y códigos geográficos del Sistema Estadístico Nacional.

[22] I.T. Jolliffe. 2006. *Principal Component Analysis.* Springer New York.

[23] Sabrina Juran, P Niclas Broer, Stefanie J Klug, Rachel C Snow, Emelda A Okiro, Paul O Ouma, Robert W Snow, Andrew J Tatem, John G Meara, and Victor A Alegana. 2018. Geospatial mapping of access to timely essential surgery in sub-Saharan Africa. *BMJ global health* 3, 4 (2018), e000875.

[24] Chaogui Kang, Song Gao, Xing Lin, Yu Xiao, Yihong Yuan, Yu Liu, and Xiujun Ma. 2010. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Geoinformatics, 2010 18th International Conference on*. IEEE, 1–7.

[25] Charles Kooperberg and Charles J Stone. 1991. A study of logspline density estimation. *Computational Statistics and Data Analysis* 12, 3 (1991), 327–347.

[26] OpenStreetMap contributors. 2018. Retrieved from https://planet.osm.org.

[27] Annette Prüss-Üstün and Maria Neira. 2016. *Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks.* World Health Organization.

[28] Carlos Sarraute, Carolina Lang, Juan de Monasterio, and Diego Weinberg. 2015. Descubriendo Chagas con Big Data. In *XVII Simposio Internacional sobre Enfermedades Desatendidas*.

[29] S. Solomon and S. Sawilowsky. 2009. Impact of Rank-Based Normalizing Transformations on the Accuracy of Test Scores. *Journal of Modern Applied Statistical Methods* (2009).

[30] Judith Timyan, Griffey Brechin SJ, Diana M Measham, and Bisi Ogunleye. 1993. Access to care: more than a problem of distance.

[31] Zhenzhen Wang, Sylvia Y He, and Yee Leung. 2018. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society* 11 (2018), 141–155.

DL4G-SDE 2019 Chairs' Welcome

It is our great pleasure to welcome you to the 2019 International Workshop on Deep Learning for Graphs and Structured Data Embedding (DL4G-SDE). The workshop provides a forum for presenting the most recent advances in embedding and representation learning for structured data as well as deep learning for graphs to unearth rich knowledge.

The call for papers attracted submissions from the United States, Europe, Asia and Australia. Each submission was reviewed by two to three reviewers. After a rigorous reviewing process, we were able to accept 9 papers (4 as regular papers for oral presentation and 5 as poster presentations that cover ongoing or recently-published work). In addition, the workshop features three keynotes delivered by leading experts in representation learning on graphs and structured data from both academia and industry:

- Philippe Cudré-Mauroux, University of Fribourg, Switzerland
- Jure Leskovec, Stanford University, USA
- Hao Ma, Facebook Artificial Intelligence Research

We believe that the program provides a good balance between trending topics in the emerging fields of deep learning on graphs and structured data embedding.

We would like to thank all participants of the workshop for making it happen, as well as The Web Conference 2019 organizers and ACM for their support. We hope that you will find the program attractive and thought provoking and that the workshop will provide you with a valuable opportunity to share ideas with other researchers and practitioners from institutions around the world. We are looking forward to a very exciting and rewarding workshop.

Jie Tang
Co-Chair
Tsinghua University

Hui Xiong
Co-Chair
Rutgers University

Michalis Vazirgiannis
Co-Chair
École Polytechnique

Ling Chen
Co-Chair
University of Technology Sydney

Yuxiao Dong
Co-Chair
Microsoft Research

Bin Li
Co-Chair
Fudan University

Fragkiskos Malliaros
Co-Chair
CentraleSupélec

Lu Qin
Co-Chair
University of Technology Sydney

THE WEB CONFERENCE

DL4G-SDE 2019 Organization

Workshop/Track Chairs:
Ling Chen (University of Technology Sydney)
Yuxiao Dong (Microsoft Research)
Bin Li (Fudan University)
Fragkiskos Malliaros (CentraleSupélec)
Lu Qin (University of Technology Sydney)
Jie Tang *(Tsinghua University)*
Michalis Vazirgiannis *(École Polytechnique)*
Hui Xiong (Rutgers University)

Program Committee:
Christos Giatsidis (École Polytechnique)
Sujuan Hou (University of Technology Sydney)
Ralf Krestel (Hasso Plattner Institute, University of Potsdam)
Viad Niculae (Instituto de Telecomunicações, Lisbon)
Giannis Nikolentzos (École Polytechnique)
Guansong Pang (The University of Adelaide)
Kevin Scaman (Huawei Technologies, Paris)
Kaisong Song (Alibaba Group)
Wei Wu (University of Technology Sydney)

THE WEB CONFERENCE

Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain

Oghenemaro Anuyah, Ion Madrazo Azpiazu, Maria Soledad Pera
People and Information Research Team (PIReT)
Department of Computer Science
Boise State University
Boise, Idaho, USA
{oghenemaroanuyah,ionmadrazo}@u.boisestate.edu,solepera@boisestate.edu

ABSTRACT

To capitalize on the benefits associated with word embeddings, researchers working with data from domains such as medicine, sentiment analysis, or finance, have dedicated efforts to either taking advantage of popular, general-purpose embedding-learning strategies, such as Word2Vec, or developing new ones that explicitly consider domain knowledge in order to generate new domain-specific embeddings. In this manuscript, we instead propose a *mixed strategy* to generate enriched embeddings specifically designed for the *educational* domain. We do so by leveraging FastText embeddings pre-trained using Wikipedia, in addition to established educational standards that serve as structured knowledge sources to identify terms, topics, and subjects for each school grade. The results of an initial empirical analysis reveal that the proposed embedding-learning strategy, which infuses limited structured knowledge currently available for education into pre-trained embeddings, can better capture relationships and proximity among education-related terminology. Further, these results demonstrate the advantages of using domain-specific embeddings over general-purpose counterparts for capturing information that pertains to the educational area, along with potential applicability implications when it comes to text processing and analysis for K–12 curriculum-related tasks.

CCS CONCEPTS

• **Information systems** → **Document representation**; Information retrieval; Data encoding and canonicalization; • **Computing methodologies** → *Knowledge representation and reasoning*; Lexical semantics.

KEYWORDS

word embeddings, domain knowledge, text representation, enriched embeddings, K-12 curriculum, structured knowledge

ACM Reference Format:
Oghenemaro Anuyah, Ion Madrazo Azpiazu, Maria Soledad Pera. 2019. Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316583

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316583

1 INTRODUCTION

Word-embeddings have become a de-facto standard for vocabulary representation. This is due to their ability to capture word contexts and varied word relations in a numerical fashion, allowing for effective and efficient word operations via their vectors. Moreover, word embeddings has been used with great success in areas related to text-processing, given their potential to achieve good generalization [11, 29].

Among the most common strategies to generate numerical word representations we find those proposed by Mikolov et al. [21], Bojanowski et al. [2], and Pennington et al. [23]. Pre-trained embeddings created as a result of using the aforementioned strategies on well-known general-purpose corpora, including Wikipedia and GoogleNews corpus, have been widely adopted by the research community as the starting point for a broad variety of applications, from analyzing reviews or scholarly articles to enhance the recommendation process [5, 14] to examining student responses in the K–12[1] context [4] and leveraging pre-trained embeddings in order to learn bilingual word representations [1].

More often than not, embeddings pre-trained using domain-agnostic resources effectively capture word co-occurrences, as well as other semantic relationships among words. Unfortunately, relationship constraints that are inherent of domain-specific contexts may be overlooked. Khatua et al. [16] echo this observation and highlight that for domains such as biomedicine, a general-purpose corpora might not yield word embeddings that best represent the domain, no matter how large that corpora is. Instead, Khatua et al. [16], along with other authors [6, 22], argue in favor of relying on existing domain-dependent corpora, such as PubMed (for medicine-related tasks) or Geological Society and Norwegian Petroleum Directory (for oil and gas domain), in order to generate more representative embeddings, while still relying on popular modeling techniques, such as Word2Vec [21] or FastText [2], to create them. Domain-dependent corpora, however, might not be as abundant or rich in resources as general-purpose counterparts.

In order to take advantage of both large general-purpose corpora and domain-specific resources, other researchers instead bias known deep learning modeling strategies towards specific contexts by altering loss functions. Such is the case of the work introduced in [13, 24], which explicitly consider the polarity of terms in creating

[1] K–12 is a common American expression that refers to primary and secondary schooling, from Kindergarten to the 12^{th} grade.

new word embeddings—doing so is a must if these embeddings are to be used for applications where the sentiment connotation can greatly influence their outcomes, such as review analysis, recommendation systems, and natural language understanding.

Beyond health, medicine, and sentiment analysis, one domain that could benefit from the availability of domain-dependent embeddings is (primary and secondary) *education*. Aligning (web) resources to K–12 curriculum delimitations to address fundamental tasks, such as finding similar exercises [17, 25] or aiding teachers with instruction design [10, 28], requires deep understanding of domain-specific terminology and the relationship among these terms. Knowledge-bases or ontologies could serve as guidelines for identifying connections across terms particular to a domain. Yet, to the best of our understanding, such information sources are not available for our area of interest; and even if they did, jargon specific to that area would not necessarily be properly captured in the vector space.

In this manuscript, we introduce Edu2Vec, a novel *mixed strategy* to generate embeddings that are meaningful to the education domain and that can serve as the foundation for solutions to the aforementioned problems in the classroom setting. Edu2Vec leverages (i) the skip-gram model presented by Mikolov et al. [21], a well-known, state-of-the-art word modeling architecture, (ii) embeddings pre-trained using Wikipedia, (iii) Wikibooks, a rich and publicly-available corpus, and (iv) structured knowledge from educational standards.[2] Edu2Vec preserves the quality of original word embeddings while requiring minimal computation for generating representations for educational concepts, by avoiding to retrain word embeddings. Furthermore, Edu2Vec is able to generate a joint space that shares both word embeddings and representations of educational concepts, which has the potential to enable multiple educational domain applications.

We conducted initial experiments in order to demonstrate the advantages of our mixed strategy. Outcomes from these experiments, along with a preliminary qualitative assessment yielded promising results in terms of the validity of infusing minimal structured area knowledge into an existing word embeddings space for learning specific representations pertaining to the K–12 curriculum.

The remaining of this paper is organized as follows. In order to contextualize our work, we discuss background information and related literature in Section 2. Thereafter, in Section 3, we present our mixed strategy for aligning structured domain knowledge with existing word embeddings in order to better capture the connections among terms in an educational context. We discuss the results of the experiments conducted to verify the validity of our proposed strategy in Section 4. Lastly, we offer some concluding remarks and directions for future work in Section 5.

2 BACKGROUND & RELATED WORK

In this section, we briefly introduce the concept of word embeddings. Thereafter, we discuss strategies that can be used to influence embedding distributions in the space as well as literature pertaining to the generation of domain-specific word embeddings.

[2]We shared both the educational hierarchy and the embedding representations for each concept at https://bit.ly/2EGHV5O

2.1 Word Embeddings

A *word embedding* is a numerical representation of a word in a multidimensional space [21]. Each dimension represents a latent feature which partially describes the meaning of a word, facilitating the comparison, in terms of meaning, of any word pair in a numerical way. Word embeddings have been used as part of solutions to problems in multiple text-related areas of study, including Recommendation Systems [30], Information Retrieval [12, 20] and Natural Language Processing [9, 33]. Key aspects that contribute to the recent success of word embeddings are (1) their ease to both be trained and used for different languages, as they just need an unlabeled corpora for doing so, and (2) their generalization capabilities, as models that are trained on small amounts of data are able to recognize words that were never used as part of the training process. Among the most popular embedding generation strategies we find: (i) CBOW and skip-gram models presented by Mikolov et al. [21] (Word2Vec), which are based on a neural model for learning representations, (ii) the model by Bojanowski et al. [2] (FastText), which leverages sub-word information in order to improve upon Mikolov's model, and (iii) the model presented by Pennington et al. [23] (GloVe), which relies on global word-word co-occurrence statistics from a corpus in order to compute word representations.

2.2 Domain-specific Embeddings

Most of the current strategies for learning word representations rely on the distributional aspect of words among a given corpora, creating word representations that are more similar to each other the more times two words co-occur in a context window. This poses a problem for specific domains, given that co-occurrence might not always imply similarity. Sentiment analysis is an example of such domain. As mentioned by Tang et al. [27], the words *good* and *bad* tend to be located close to each other in the embedding space given that they often appear in similar contexts. Unfortunately, this spatial distribution is counterproductive for sentiment analysis, given the opposite polarity of the these two terms. To address this limitation, Tang et al. [27] explicitly encode sentiment information into the continuous representations of words. Piao and Breslin [24] focus on incorporating sentiment prediction information into the general loss function, whereas Hamilton et al. [13] combine both semantic high-quality word-embeddings with an sentiment label propagation approach.

Other domains that benefit from domain-specific embeddings are health, medicine, finance, current events, and geosciences, to name a few. In their cases, the use of domain-specific corpora of diverse sizes to learn embeddings more suitably distributed in the space is the most prominent strategy [6, 16, 18]. Alternatively, researchers have also considered leveraging knowledge resources (and in some cases multiple objective functions) in order to generate enhanced embeddings that capture additional semantic relationships, such as hyponymy and relatedness in the target domain [22, 26, 32].

Research efforts dedicated to tackle the issue of generating domain-specific representations are a must if we want to advance applications that need to process texts that incorporate specialized jargon. Hierarchical representations are a natural way to represent human understanding of domain-specific concepts, which is why we believe it is possible to leverage such structures for computing domain

dependent concept representations, with minimal computational time, as we describe in the following Section.

3 METHODOLOGY

In this section, we describe the strategy we propose for incorporating structured knowledge from the educational domain into the word embedding space. The proposed strategy consists of two steps: (i) building the educational concept hierarchy and (ii) aligning hierarchy concepts with existing, pre-trained word embeddings.

3.1 Building the Educational Hierarchy

We first build a hierarchy that captures educational knowledge. This structured knowledge, which resembles the tree structure in Figure 1(a), aligns *subjects* that pertain to the K–12 curriculum to different *grade levels*. In this structure, we also capture *topics* that correspond to the different subjects. For instance, the topics *geometry* and *chemical reactions* are aligned to the *Mathematics* and *Science* subjects, respectively. Furthermore, each of the topics in our hierarchy is connected to a number of *words*, which are terms often associated with these topics. For example, *triangles* is a term that educators expect their students to be familiar with for the *geometry* topic, whereas *molecule* is part of the vocabulary commonly affiliated with the *chemical reactions* topic. (See Figure 1(b) for another example of structured knowledge in the K–12 curriculum).

In creating our knowledge structure, we leverage several well-established *Educational Standards* that serve as guidelines to identify essential information about the K–12 curriculum, i.e., subjects, topics, and relevant vocabulary across different school grade levels:

CCSS – Common Core State Standards Established in 2009, the common core is a set of college and career-ready standards for children from Kindergarten through 12^{th} grade in English and Mathematics [7]. Some English topics include *Reading and literacy* and *Persuasive writing*, while some Mathematics topics are *Geometry*, as well as *Data Analysis, Probability, and Statistics*.

NGSS – Next Generation Science Standards The NGSS are science content standards that set expectations of what children should know in science related subjects such as Chemistry, Biology, Geography, and Solar System [8]. The NGSS enables teachers the flexibility to design classroom experiences that enhance children's interests in Science. Some NGSS topics include *Structure and Properties of Matter* and *Earth's Systems*.

ICS – Idaho Content Standards ICS details what children in Idaho public schools ought to know at the end of each grade in subjects like Sciences, Social Studies, Mathematics, and English (ICS adopts CCSS guidelines for Mathematics and English) [15]. We depend on ICS to identify topics relating to children's Social Studies, Mathematics, and English subjects.

Based on these standards, we identify four subjects: Mathematics, English, Social Studies, and Science[3]. For each of these subjects, we automatically extract educational topics using a parser that scans through the written guidelines provided by CCSS, NGSS, and ICS. A summary of the number of extracted educational topics that map to each of the subjects in our hierarchy is presented in Table 1.

Table 1: Topics and Wikibooks articles that correspond to each subject in our educational hierarchy.

Subject	Number of Topics	WikiBooks Articles
Mathematics	7	1,441
English	8	7,392
Science	40	4,500
Social Studies	27	1,645
Total	**82**	14,978[4]

In order to facilitate mapping educational concepts in the hierarchy (i.e., subjects and topics) to an existing general-purpose word embeddings (as we discuss in Section 3.2), we require the existence of *representative words* that correspond with the aforementioned educational concepts. In our case, we explore topic descriptions available on CCSS, NGSS, and ICS and identify the top–20[5] most representative words for each topic, based on TF-IDF scores. We are aware that twenty keywords may not be sufficient to capture the breath of vocabulary commonly associated with each topic. For this reason, we turn to an external, publicly available resource that can provide known relationships among educational topics and words: *Wikibooks*[6].

Wikibooks is a Wikipedia community hosted for the purpose of creating a free library of diverse textbooks. Wikibooks articles are assigned categories in order to group articles that are related to similar subjects. For instance, Wikibooks articles on *Atmospheric Layers*, *Metamorphic Rocks*, and *Effects of Air Pollution* are grouped under the category *High School Earth Science*. We argue that there is a natural connection among categories used by Wikibooks and the topics defined by Educational Standards and rely on such alignment in order to identify the articles within Wikibooks that pertain to educational concepts discussed in the K–12 curriculum. We do so using Algorithm 1, which iterates through Wikibooks articles and the 82 topics in our hierarchy and selects a Wikibooks article if its category contains one of the topics as a substring[7]. Using the resulting set of 13,543 Wikibooks articles matching the K–12 curriculum, we once again identify the top–20 most representative terms that describe each article[8] and link those with each corresponding educational topic.

Based on the grade, subject, topic, and words pairs gathered using both educational standards and Wikibooks, we use a number of relationship types to create <grade, subject>; <subject, topic>; and <topic, word> pairs. Overall, we identify 86,295 relation pairs (described below) that we use to build our educational hierarchy from domain knowledge.

[3]Note that we use Science to broadly capture subjects such as Chemistry, Physics, Geography, and Biology.

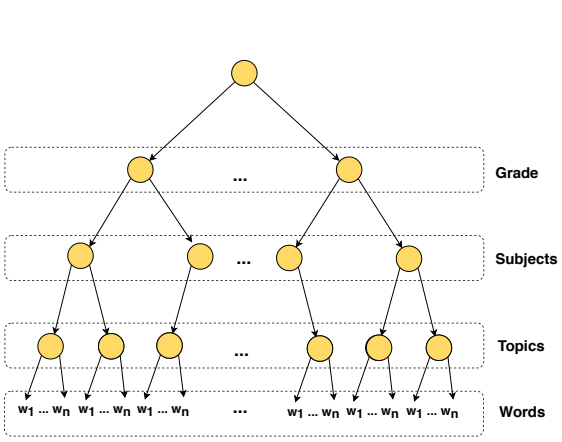
[4]This number differs from the ideal 13,543 articles being that we found a number of overlaps in Wikibooks articles for the different subjects, as some articles have multiple categories, making them align with more than one subject in some cases.

[5]We set this parameter value empirically.

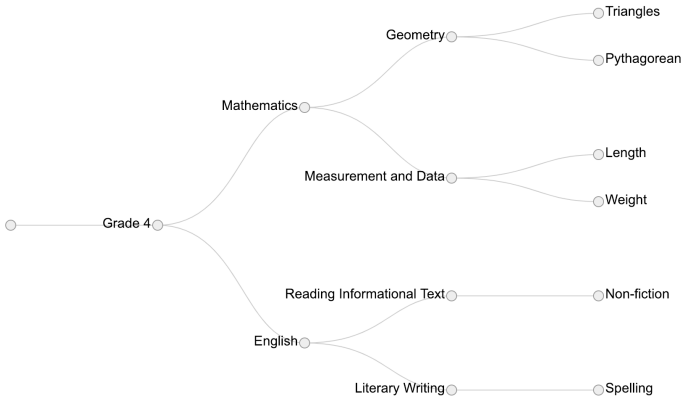
[6]https://en.wikibooks.org

[7]Tokens in topics and categories may vary as a result of plurals, verb tenses, or suffixation. Hence, we first stem each of their respective tokens using the Porter stemmer algorithm [31].

[8]We identify representative terms using TD-IDF scores; with IDF calculated based on all Wikibooks articles, not just the ones matching K–12 curriculum



(a) Structured domain knowledge.



(b) Example nodes for the 4^{th} grade.

Figure 1: Education domain hierarchy, including grades, subjects, topics, and term relations.

Grade-Subject relations. We identify several relationships between grade levels and subjects by using the guidelines provided by the aforementioned educational standards. In total, our hierarchy consists of thirteen grade levels (K–12) and four subjects.

Subject-Topic relations We create relationships that capture the explicit connections among subjects and topics. This yields a hierarchy comprised of 82 unique topics for the subjects examined.

Topic-Word relations For each leaf topic in our hierarchy we match them to conceptual words, which we identify using both the educational standards and Wikibooks. Through this process, we identified a total of 15,420 unique terms for the educational topics.

Algorithm 1 Map Wikibooks articles to K–12 topics

```
Input: T, a list of K--12 Topics, WP a collection of Wikibooks articles
Output: WPT a subset of articles from WP that map to topics in T
WPT = empty set
for each t in T do
    for each w in WP do
        page_cat = extractCategory(w)
        if all(stemmed (t)) in stemmed(page_cat) then
            Add <t, w> to WPT
        end if
    end for
end for
return WPT
```

3.2 Aligning Structured Knowledge and the Word Space

In designing our mixed strategy, we aim to leverage existing general-purpose resources, i.e., pre-trained embeddings, in addition to domain-specific resources, i.e., the educational hierarchy introduced in Section 3.1. We define the latter as $H = (E, V)$, where E is the set of edges in the hierarchy and V the set of vertices. H is composed by two disjoint sub-graphs H_c and H_w that contain educational concepts (grades, subjects, topics) and words, respectively. Moreover, we define the vocabulary considered by Edu2Vec as D, which is comprised of the set of words D_w (from pre-trained word embedding vocabulary) and the set of educational concepts D_c that correspond to each vertex in H_c; $D = D_w \bigcup D_c$ and $|D_w \bigcap D_c| = 0$.

Let $X \in \mathbb{R}^{|D| \times 300}$ be a two-dimensional vector containing all the embeddings used by Edu2Vec, where the i^{th} row of X corresponds to the i^{th} term in D, $|D|$ is size of D (i.e., the number of distinct vocabulary terms) and 300 is the number of latent features used for representing each embedding. X results from stacking two vectors $X_w \in \mathbb{R}^{|D_w| \times 300}$ and $X_c \in \mathbb{R}^{|D_c| \times 300}$; the former represents the embeddings for words in D_w and the latter educational concepts in V_c. In our case, we take advantage of word embeddings pre-trained using FastText[9] in Wikipedia for initializing X_w. As for X_c, we initialize it using a random uniform distribution in range $[-1, 1]$.

In order to generate representations of educational concepts related to the K-12 curriculum that are aligned with the word representation space, we extend the model presented by Mikolov et al. [21], enabling it to consider a tree structure as input. This strategy, depicted in Figure 2, allows us to explicitly consider the domain-specific hierarchical structure H presented in Section 3.1 for learning educational concept representations X_c for H_c.

In aligning embeddings in X_c with those X_w, we maximize the following function:

[9] https://fasttext.cc/

$$\max_{X_c} \sum_{(v,v')\in V\times V} \frac{1}{d(v,v')} * P(v \mid v') \quad (1)$$

where $d(v, v')$ represents the distance between concept vertex v and v', which is used to attribute less importance to nodes that are further apart, and the probability function P is modeled as:

$$P(v \mid v') = \frac{\exp(X_v{}^\top X_{v'})}{\sum_{i=1}^{|D|} \exp(X_i{}^\top X_{v'})} \quad (2)$$

where X_v and X'_v are the embedding representations of v and v' and X_i is the embedding representation of the i^{th} token in D.

We use stochastic gradient descend [3] with a learning rate of 0.01 for minimizing the following function:

$$\min_{X_c} - \sum_{(v,v')\in V\times V} \frac{1}{d(v,v')} * P(v \mid v') \quad (3)$$

It is important to note that we only minimize in terms of X_c, excluding any update to X_w induced by the calculated gradient. This ensures that the pre-trained word embeddings are never changed, which guarantees the quality of the aforementioned embeddings remains intact, i.e., it is not affected by the minimization introduced by Edu2Vec. The outcome of the optimization procedure is a trained X_c which contains the representations of all concepts in the educational hierarchy H_c. These representations are aligned with pre-trained word representations X_w in a joint space. A joint space of both educational concepts and word embeddings is what makes possible to directly match keywords (in any text) with respect to K–12 curriculum terminology, which can be exploited in the development of educational text processing applications.[10]

4 EXPERIMENTAL RESULTS

We offer below insights that result from qualitative and quantitative experiments conducted to demonstrate the validity of our newly-learned educational representations.

4.1 Subject - Topic Similarity Task

To demonstrate the applicability of our domain-specific strategy, we first measure the degree of closeness for each of the 82 < *subject*, *topic* > pairs extracted from the Education Standards defined in Section 3.1. We do so by computing the cosine similarity between the keywords of each pair using the embeddings generated by Edu2Vec. To contextualize the results, we also compute the cosine similarity using general-purposed embeddings. Specifically, we use FastText word vectors for the English language trained on Wikipedia.[11] These vectors with a dimension of 300 were generated using the skip-gram model with default parameters. It is important to note that not all topics are unigrams; however, FastText's dictionary is comprised of single keywords, not n-gram phrases. For this reason, we handle topics that are n-grams phrases (for $n > 1$) by splitting the topic words into individual tokens before computing the cosine similarity between the embeddings of the < *subject*, *topic* > pairs.

We present in Figure 3 the similarity distribution computed using Edu2Vec and FastText. We see that the use of Edu2Vec leads to higher similarity scores for the majority of the pairs than when using general-purpose word embeddings. We infer from this distribution that indeed, a strategy that explicitly models structured knowledge from a given domain, i.e., education in our case, is able to better capture connections among higher-order level concepts (e.g., subjects and topics) than word-level embeddings trained with a general corpora. To present a fair picture on the different similarity distributions, we also include in Figure 3 the values computed using FastText representations only for topics that are originally unigrams. Even then, most of the yielded scores are in the [0, 0.4] range. We argue that these results serve as evidence of the benefits of integrating information from the educational domain, even if this information is limited in size, to enrich embeddings so that they can be leveraged by applications pertaining to text analysis in the corresponding domain.

4.2 Topic - Term Similarity Task

To further demonstrate the need for creating domain-specific concept representations, we also explore connections among topics and terms in the educational domain. For our analysis, we adapt the evaluation framework presented in [32], which is centered on exploring how a traditional skip-gram model and other knowledge-enriched strategies perform in identifying the topic that is the most related to a given term.

Due to the lack of existing benchmarks and datasets, we create our own using *SimpleWiki* pages[12]. We turn to SimpleWiki as it is a publicly available resource and, more importantly, it is an edition of Wikipedia that is oriented to and suitable for young audiences which uses Wikipedia's page categorization system. Much like we described in Section 3.1 for WikiBooks, we take advantage of categories assigned to each SimpleWiki page and create two datasets, which we use as ground truth. In the first dataset, DS_A, we map educational topics to high level terms associated with them. To do so, we select the subset of topics in our hierarchy that exactly match categories in SimpleWiki and extract all their respective subcategories. For example, the mathematics topic *geometry* is associated with subcategories *angles*, *conic sections*, *shapes*, and *fractals*. Using this categorization, we created 463 <*topic*, *term*, *score*> tuples, where *score* is defined as 1 being that these pairs are expected to have the optimum similarity value as subcategories for their respective SimpleWiki pages have been defined by experts. In our second dataset, DS_B, we map educational topics to more general terms commonly used within the educational domain. To create DS_B, we follow the same process outlined for DS_A but instead extracting subcategories, for each SimpleWiki category (i.e., topic) we extract the titles of pages that correspond to said category. Based on this approach, we created 2,672 <*topic*, *term*, *score*> tuples; we set *score* to 1 to capture an "ideal" similarity.

We emulate the evaluation framework in [32] and measure the degree to which general-purpose and domain-specific word representations can be effectively used to map K–12 vocabulary to educational topics (i.e., topics). To do so, we use error rate as our metric, which we define in Equation 4.

[10] We are aware that some topic or subject words may overlap with words in FastText. To address this limitation, we added a prefix "T_" to terms that are also topics and "S_" to terms that are subject words.

[11] https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.en.zip

[12] https://simple.wikipedia.org/wiki/Category:

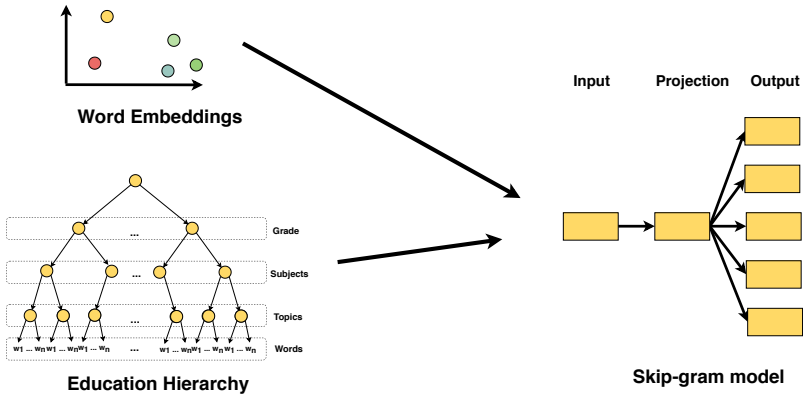


Figure 2: Proposed strategy; skip-gram model refers to version adapted based on Equations 1-3.

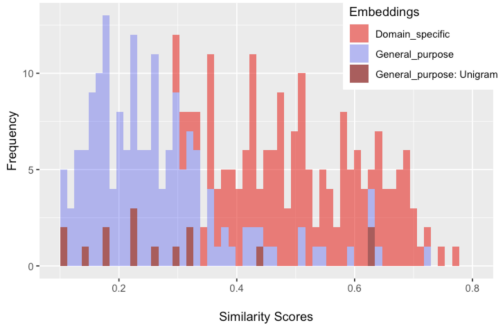


Figure 3: Similarity distribution for <subject, topic> pairs based on general-purpose (FastText) and domain-specific (Edu2Vec) strategies to create word representations.

$$Error_rate = \frac{\sum_{i=1}^{n} Sim(topic_i, term_i) - \hat{Sim}(topic_i, term_i)}{n} \quad (4)$$

where n is the number of instances in a given dataset, $Sim(topic_i, term_i)$ is the expected similarity score for $< term, topic >$ pairs in the i^{th} instance in the dataset (1 in our case), and $\hat{Sim}(topic_i, term_i)$ is the similarity score of $< term, topic >$ predicted using Equation 5.

$$\hat{Sim}(topic_i, term_i) = \frac{emb(term_i) \cdot emb(topic_i)}{\|emb(term_i)\| \cdot \|emb(topic_i)\|} \quad (5)$$

where *emb* is used to denote the embedding generated for a given term or topic.

Task A: Mapping topics to high level terms. To conduct this analysis, we use DS_A and compute the error rate using Equation 4, based on similarity scores predicted by using general-purpose and educational embeddings, generated using FastText and Edu2Vec.

As reported in Table 2, the error rates yielded using FastText's embeddings were higher than when using Edu2Vec's embeddings. We verified this difference using a statistical significance test (T-test; $p < 0.05$). This result was not unexpected, as FastText's embeddings are based on diverse, non-education-specific information and thus are not meant to explicitly capture within the vector space connections among higher-level, conceptual terminology that naturally aligns with the K–12 curriculum.

Table 2: Error rates for the topic-term similarity task using domain-specific and domain-agnostic embeddings. "[]", indicates the variance among computed values.

Experiment	Error Rate (%) Edu2Vec	FastText
Task A	**15.1** [0.0036]	61.8 [0.0338]
Task B	**16.8** [0.0056]	69.4 [0.0316]

Task B: Mapping topics to lower level terms. For this experiment, we rely on instances in DS_B, and also measure the error rates when predicting the similarity between keywords in each <*topic*, *term*> pair, using embeddings from FastText and Edu2Vec to represent topic and term. Similar to results reported for Task A, we see that using embeddings from FastText for representing topics an terms led to higher error rates when compared to Edu2Vec's embeddings. Differences in error rates are statistically significant (T-test; $p < 0.05$). We observed a 11% increase in Edu2Vec's errors from Task A. We hypothesize this is due to the fact that it is more challenging for Edu2Vec to map more general terms to higher-level concepts, i.e., Edu2Vec is trained to explicitly account for connections among educational concepts (i.e., broader, higher level terms such as subjects and topics). For instance, it is a straightforward task for Edu2Vec to situate in close proximity the topic *shapes* and the topic *geometry*, however, mapping *shapes* to the term *vertex* is more challenging, as those two keywords are presented at different abstraction levels, one is a topic, while the other is a term.

4.3 Discussion

In the previous section, we examined the effectiveness of using Edu2Vec in order to create embeddings that can capture relationships among educational subjects, topics, and terms in the vector space. To further contextualize the applicability of our initial findings, we use a number of illustrative examples. As shown in Figure 4, topics aligned to specific subjects are in close proximity

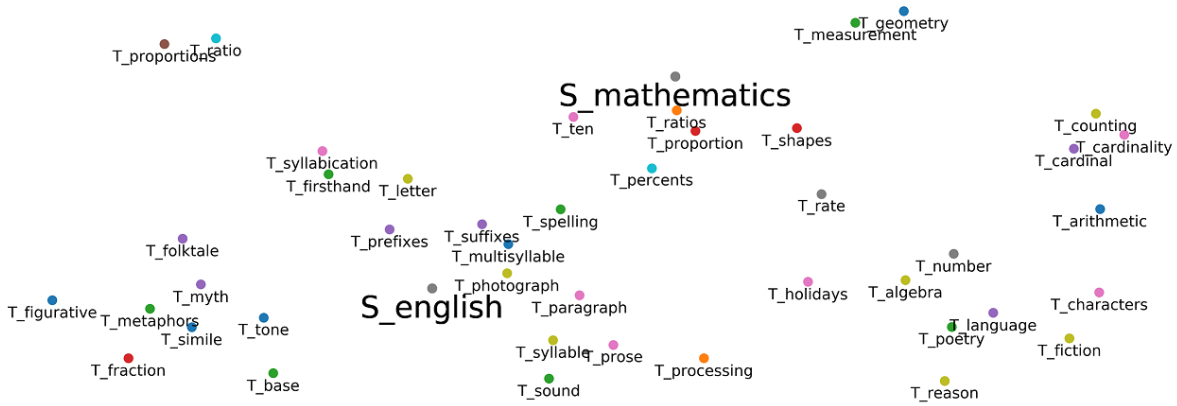


Figure 4: Space representation of subject and topic embeddings. Area around subjects *Mathematics* and *English* is shown. Dimensionality reduced using t-distributed Stochastic Neighbor Embedding (TSNE) [19] algorithm with perplexity and iteration number set to 5 and 5000 respectively.

among each other, when compared to other topics. For instance, the subject ***Mathematics*** is close to topics ***ratios***, ***proportions***, ***geometry***, and ***measurement*** but further apart from topics related to the subject ***English***, such as ***myth*** or **spelling**.

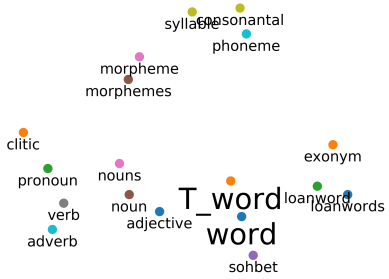


Figure 5: Closest embeddings for *word* as a topic and a term generated using Edu2Vec and FastText, respectively. Dimensionality reduced using t-distributed Stochastic Neighbor Embedding (TSNE) [19] algorithm with perplexity and iteration number set to 5 and 5000 respectively.

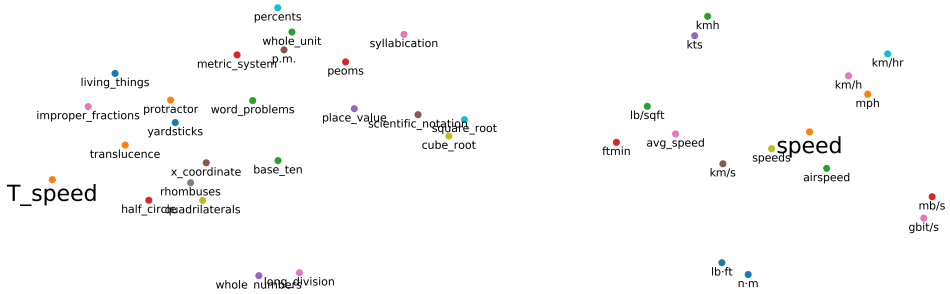
Through further manual analysis, we found that embeddings from FastText correctly capture connections among more general terms in the educational concept vector space. This is showcased in Figure 5, where we depict the embedding representation for both the educational concept ***word*** (i.e., T_word) and the term itself. In this case, there is no noticeable difference that arises by considering the word itself or the topic, in terms of neighbouring vocabulary. However, not biasing the model to favor information specific to a domain (educational in our case) may be a detriment, as then the model would not be able to identify terminology that is essential for the domain. This is indeed apparent in Figure 6, where we depict the most similar words for both the topic *T_speed* and the term *speed*. On the one hand, the term *speed* is most similar to other terms that are known to often co-occur with the keyword speed in general documents, such as *km/s* or *avg_speed*. On the other hand, the topic *T_speed* is closer to other keywords like *protractor* or *x_coordinate* that are known to appear in educational materials that focus on physical science and integrated math curriculum.

5 CONCLUSIONS & FUTURE WORK

Even if word-representation techniques are abundant, most of them are currently focused on the distributional aspect of the words. These strategies tend to generate domain-agnostic representations that can overlook important information that could greatly influence task performance in specific domains. In order to address this issue, we have introduced a novel strategy for generating representations for educational concepts. For doing so, we introduce Edu2Vec, which takes advantage of structured educational knowledge in the form of an hierarchy consisting of K–12 related information in addition to traditional pre-trained word-embeddings.

We conducted an initial assessment using labeled educational resources as well as SimpleWikipedia, demonstrating the importance of incorporating domain knowledge into the embedding-generation process, even if that knowledge is limited. These results, compounded with a preliminary qualitative analysis lead us to argue in favor of relying on embeddings generated by Edu2Vec, as opposed to traditional, general-purpose embeddings, when it comes to tasks and applications essential to the K–12 curriculum, including, but not limited to measuring the educational value of documents; filtering and ranking educational resources in response to web search tasks conducted in the classroom setting and identifying suitable documents that match pre-defined learning objectives.

In the future, we will dedicate research efforts to investigate the applicability of the proposed strategy in domains beyond Education—provided the availability of a structured domain-specific hierarchy, or the possibility to infer one from data. Furthermore, we plan on exploring strategies that can better utilize the hierarchical structure of a domain knowledge. Finally, we will consider alternative techniques for better representing educational documents incorporating not only topical information but also other aspects, such as text complexity.



(a) Closest embeddings for the ***topic*** speed (T_Speed).

(b) Closest embeddings for the ***term*** speed.

Figure 6: Closest embeddings for *speed* as a topic and a term generated using Edu2Vec and FastText, respectively. Dimensionality reduced using t-distributed Stochastic Neighbor Embedding (TSNE) [19] algorithm with perplexity and iteration number set to 5 and 5000 respectively.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation Award Number 1565937.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 451–462.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[3] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[4] Florin Adrian Bulgarov and Rodney Nielsen. 2018. Proposition entailment in educational applications using deep neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[5] Rose Catherine and William Cohen. 2017. Transnets: Learning to transform for recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 288–296.

[6] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 166–174.

[7] Common Core State Standards Initiative. Accessed: March 2018. Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects. http://www.corestandards.org/assets/CCSSI_ELA20Standards.pdf.

[8] National Research Council et al. 2014. *Developing assessments for the next generation science standards*. National Academies Press.

[9] Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 69–78.

[10] Michael D Ekstrand, Ion Madrazo Azpiazu, Katherine Landau Wright, and Maria Soledad Pera. 2018. Retrieving and Recommending for the Classroom. *ComplexRec 2018* 6 (2018), 14.

[11] Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth JF Jones. 2015. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 795–798.

[12] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 38.

[13] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, Vol. 2016. NIH Public Access, 595.

[14] Hebatallah A Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, and Alessandro Micarelli. 2018. Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 465–469.

[15] Idaho State Department of Education. Accessed: March 2018. Idaho Content Standards. https://www.sde.idaho.gov/academic/standards/.

[16] Aparup Khatua, Apalak Khatua, and Erik Cambria. 2019. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing & Management* 56, 1 (2019), 247–257.

[17] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. 2018. Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1821–1830.

[18] Yue Liu, Tao Ge, Kusum S Mathews, Heng Ji, and Deborah L McGuinness. 2018. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. *arXiv preprint arXiv:1804.04225* (2018).

[19] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[20] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. 2018. Looking for the Movie Seven or Sven from the Movie Frozen?: A Multi-perspective Strategy for Recommending Queries for Children. In *Proceedings of the 2018 Conference on Human Information Interaction& Retrieval*. ACM, 92–101.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[22] Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of Domain-specific Word Embeddings using Knowledge Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[24] Guangyuan Piao and John G Breslin. 2018. Financial aspect and sentiment predictions with deep neural networks: an ensemble approach. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1973–1977.

[25] Jiří Rihák and Radek Pelánek. 2017. Measuring Similarity of Educational Items Using Data on Learners' Performance. *Educational Data Mining* (2017).

[26] Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. 2015. Exploring word embedding for drug name recognition. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. 64–72.

[27] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1555–1565.

[28] Khoi-Nguyen Tran, Jey Han, Danish Contractor, Utkarsh Gupta, Bikram Sengupta, Christopher J Butler, and Mukesh Mohania. 2018. Document Chunking and Learning Objective Generation for Instruction Design. *arXiv preprint arXiv:1806.01351* (2018).

[29] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. ACM, 363–372.

[30] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1235–1244.

[31] Peter Willett. 2006. The Porter stemming algorithm: then and now. *Program* 40, 3 (2006), 219–223.

[32] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM international conference on*

conference on information and knowledge management. ACM, 1219–1228.

[33] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 3 (2018), 55–75.

Link Prediction with Mutual Attention for Text-Attributed Networks

Robin Brochier
Université de Lyon, Lyon 2, ERIC EA3083
Digital Scientific Research Technology
robin.brochier@univ-lyon2.fr

Adrien Guille
Université de Lyon, Lyon 2, ERIC EA3083
adrien.guille@univ-lyon2.fr

Julien Velcin
Université de Lyon, Lyon 2, ERIC EA3083
julien.velcin@univ-lyon2.fr

ABSTRACT

In this extended abstract, we present an algorithm that learns a similarity measure between documents from the network topology of a structured corpus. We leverage the Scaled Dot-Product Attention, a recently proposed attention mechanism, to design a mutual attention mechanism between pairs of documents. To train its parameters, we use the network links as supervision. We provide preliminary experiment results with a citation dataset on two prediction tasks, demonstrating the capacity of our model to learn a meaningful textual similarity.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; *Artificial intelligence*; • **Information systems** → *Data mining*.

KEYWORDS

representation learning; link prediction; attributed network; natural language processing

ACM Reference Format:
Robin Brochier, Adrien Guille, and Julien Velcin. 2019. Link Prediction with Mutual Attention for Text-Attributed Networks. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316587

1 RELATED WORKS

In this section, we relate recent works in the fields of network embedding (NE) and attention mechanism for natural language processing (NLP).

1.1 Attributed Network Embedding

DeepWalk [6] first proposed to derive the word embedding algorithm *Word2vec* [4] by generating paths of nodes, akin to sentences, with truncated random walks. *DeepWalk* and other variants are generalized into a common matrix factorization framework in *NetMF* [7]. To extend *DeepWalk* for text-attributed networks, *TADW* [10] expresses this latter as a matrix factorization problem and incorporates a matrix of textual features T, produced by latent semantic indexing (*LSI*), into the factorization so that the vertex similarity matrix can be reconstructed as the product of three matrices V^T, H and T.

1.2 Attention Mechanisms for NLP

The *Transformer* [9] is a novel neural architecture that outperforms state-of-the-art methods in neural machine translation (NMT) without the use of convolution nor recurrent units. The Scaled Dot-Product Attention (SDPA) is the main constituting part of the *Transformer* that actually performs attention over a set of words. It takes as input a query vector q and a set of key vectors K of dimensions d_k and value vectors V of dimensions d_v. One weight for a value is generated by a compatibility function with its corresponding key and the query. Formally, the attention vectors are generated in parallel for multiples queries Q, following the formula: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$. The result is a set of L attention vectors (L being the number of queries) of dimension d_v. The matrices Q, K and V are produced by projection of initial words representations W with three matrices P^Q, P^K and P^V whose parameters are meant to be learned.

Several works [2, 8] adapted the *Transformer* architecture beyond the task of NMT. Their main idea is to train the *Transformer* in an unsupervised fashion over large corpora of texts and further refine its parameters on a wide variety of supervised tasks. Motivated by these recent works, we present a model, *MATAN* (**M**utual **A**ttention for **Te**xt-**A**ttributed **N**etworks), that derives from the the SDPA to address the task of link prediction in a network of documents.

2 PROPOSED MODEL

We propose an algorithm for link prediction in text-attributed network. Our model is trained under a NE procedure, presented in Section 2.1. The optimization of the reconstruction error is performed via dot-product between contextual document representations e_u^v and e_v^u. These embeddings are generated with a mutual attention mechanism over their textual contents only, described in Section 2.2.

2.1 Overall Optimization

The model takes as input a network of documents $G = (V, E, T)$, T being the textual content of the documents. We precompute word embeddings W and a normalized similarity measure between nodes M designed from the adjacency matrix A of the network. Each document t_u is associated with a bag of word embeddings

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316587

W^{t_u} matrix. For any pair of node $(u, v) \in V^2$, mutual embeddings are generated with an asymmetric mutual attention function f^A_Θ for both documents given their bags of word embeddings $e^v_u = f^A_\Theta(W^{t_u}, W^{t_v})$ and $e^u_v = f^A_\Theta(W^{t_v}, W^{t_u})$. We define the unormalized similarity between the two nodes as the dot product of their mutual embeddings $e^v_u.e^u_v$. We aim at learning the parameters Θ by minimizing the KL divergence from the similarity distributions M (from the graph) to that of the normalized distribution of the dot products between the mutual embeddings (text associated to the nodes). We achieve this by employing noise-contrastive estimation, minimizing the following objective function: $J = -\sum_{(u,v)\in C}\left(\log\sigma(e^v_u \cdot e^u_v) + \sum_{i=1}^{k}\mathbb{E}_{z\sim q}\left[\log\sigma(-e^z_u \cdot e^u_z)\right]\right)$, where σ is the sigmoid function. C is a corpus of pairs of nodes generated by drawing uniformly existing links from the empirical distribution of links M. k negative nodes are uniformly drawn for each positive pair. To minimize this objective, we employ stochastic gradient descent using ADAM [3].

2.2 Mutual Attention Mechanism

The role of $f^A_\Theta(W^{t_u}, W^{t_v})$ is to generate a contextual representation of t_u given t_v. The parameters Θ we aim to learn are composed of three matrices $\Theta = \{P^Q, P^K, P^V\}$ of dimension $D\times D$ each. For all words of the target document t_u, we create queries $Q_u = W^{t_u}P^Q$. We similarly create keys and values from the contextual document t_v, such that $K_v = W^{t_v}P^Q$ and $V_v = W^{t_v}P^V$. Attention representations for each target word are then computed, following the SDPA formula: $\text{SDPA}_\Theta(W^{t_u}, W^{t_v}) = \text{softmax}(\frac{Q_u K_v^T}{\sqrt{D}})V_v$. Note that $\text{SDPA}_\Theta(W^{t_u}, W^{t_v})$ has dimension $L \times D$, that is, we have a mutual attention representation of each word of document t_u given t_v. Finally, the representation for document t_u is obtained by averaging its word mutual attention vectors: $e^v_u = f^A_\Theta(W^{t_u}, W^{t_v}) = \sum_{i=0}^{L}\text{SDPA}_\Theta(W^{t_u}, W^{t_v})_i$. Similarly, e^u_v is generated by flipping indices u and v. The intuition behind this model is that the matrices P^Q and P^K learn to project pairs of words that explain links in the network such that their dot-products produce large weights. P^V is then meant to project the word vectors such that their average produces similar representations for nodes that are close in the network and dissimilar for nodes that are far in the network.

3 EXPERIMENTS

To assess the quality of our model, we perform two tasks of link prediction on a dataset of citation links between scientific abstracts: Cora [1]. The first prediction evaluation, called edges-hidden, consists in hiding a percentage of the links given a network of documents and measuring the ability of the model to predict higher scores to hidden links than to non-existing ones by computing the ROC AUC. The second evaluation, called nodes-hidden, consists in splitting the network into two unconnected networks, keeping a percentage of the nodes in the training network.

We precompute on the full corpus word embeddings using *GloVe* [5] of dimension 256 with a co-occurrence threshold $x_{\max} = 10$, a window size $w = 5$ and 50 epochs. We precompute *LSI* [1] vectors of dimension 128. For the edge-hidden prediction task, we provide results performed by *NetMF* with $k = 10$ negative samples. *TADW* is run with 20 epochs and *MATAN* is performed with $k = 1$ negative sample and 10^5 sampled pairs of documents. The empirical similarity matrix between the nodes we chose is the normalized adjacency matrix. All produced representations are of dimension 256.

[1]Get the data: https://linqs.soe.ucsc.edu/data

3.1 Results

Table 1: Edges-hidden link prediction ROC AUC

% of training data	10%	20%	30%	40%	50%
NeMF	59.0	67.2	77.5	83.2	87.2
TADW	68.0	82.0	87.1	**93.2**	**94.5**
MATAN	**82.3**	**87.1**	**88.6**	90.9	91.0

Table 2: Nodes-hidden link prediction ROC AUC

% of training data	10%	20%	30%	40%	50%
TADW	64.2	**75.8**	**80.3**	**81.9**	**82.3**
MATAN	**69.4**	73.0	75.4	77.9	78.6

Tables 1 and 2 show the results of our experiments. *MATAN* shows promising results for learning on a small percentage of training data on both evaluations. *TADW* has better scores for nodes-hidden predictions which might be explained by the capacity of *LSI* to learn discriminant features on a small dataset unlike *GloVe*. In future work we would like to deal with bigger datasets from which word embedding methods might capture richer semantic information.

REFERENCES

[1] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41, 6 (1990), 391–407.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[6] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[7] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 459–467.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper. pdf* (2018).

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[10] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network representation learning with rich text information.. In *IJCAI*. 2111–2117.

Real-Time Streaming Graph Embedding Through Local Actions

Xi Liu
Texas A&M University
xiliu.tamu@gmail.com

Ping-Chun Hsieh
Texas A&M University
pingchun.hsieh@tamu.edu

Nick Duffield
Texas A&M University
duffieldng@tamu.edu

Rui Chen
Samsung Research America
rui.chen1@samsung.com

Muhe Xie
Samsung Research America
muhexie@gmail.com

Xidao Wen
University of Pittsburgh
xidao.wen@pitt.edu

ABSTRACT

Recently, considerable research attention has been paid to graph embedding, a popular approach to construct representations of vertices in latent space. Due to the curse of dimensionality and sparsity in graphical datasets, this approach has become indispensable for machine learning tasks over large networks. The majority of the existing literature has considered this technique under the assumption that the network is static. However, networks in many applications, including social networks, collaboration networks, and recommender systems, nodes, and edges accrue to a growing network as streaming. A small number of very recent results have addressed the problem of embedding for dynamic networks. However, they either rely on knowledge of vertex attributes, suffer high-time complexity or need to be re-trained without closed-form expression. Thus the approach of adapting the existing methods designed for static networks or dynamic networks to the streaming environment faces non-trivial technical challenges.

These challenges motivate developing new approaches to the problems of streaming graph embedding. In this paper, we propose a new framework that is able to generate latent representations for new vertices with high efficiency and low complexity under specified iteration rounds. We formulate a constrained optimization problem for the modification of the representation resulting from a stream arrival. We show this problem has no closed-form solution and instead develop an online approximation solution. Our solution follows three steps: (1) identify vertices affected by newly arrived ones, (2) generating latent features for new vertices, and (3) updating the latent features of the most affected vertices. The new representations are guaranteed to be feasible in the original constrained optimization problem. Meanwhile, the solution only brings about a small change to existing representations and only slightly changes the value of the objective function. Multi-class classification and clustering on five real-world networks demonstrate that our model can efficiently update vertex representations and simultaneously achieve comparable or even better performance compared with model retraining.

ACM Reference Format:
Xi Liu, Ping-Chun Hsieh, Nick Duffield, Rui Chen, Muhe Xie, and Xidao Wen. 2019. Real-Time Streaming Graph Embedding Through Local Actions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316585

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316585

1 INTRODUCTION

Recently graph embedding, also known as graph (a.k.a network) representation learning, has received considerable research attention. That is due to the fact that many real-world problems in complex systems can be modeled as machine learning tasks over large graphs. Direct representations of a vertex in such graphs are usually by its adjacency vector and thus suffer from the curse of dimensionality and sparsity. The idea of graph embedding is to learn a low-dimensional and dense representation for vertices in latent space. The latent representations are learned with the objective of preserving the structural information of the original graph into the geometric relationships among vertices' vector representations [1]. The learned vertex representations are regarded as informative feature inputs to various machine learning tasks. Graph embedding has been proven to be a useful tool for many machine learning tasks, such as vertex classification [2], community detection [3], and link reconstruction [4].

Prior studies have proposed several prominent graph embedding methods [2, 5–9] (see Section 2 for a careful review). Unfortunately, they are subject to three limitations. First, these methods have focused on *static* graphs. However, the majority of real-world networks are naturally dynamic and continuously growing. New vertices, as well as their partial edges, form in a streaming fashion. Such networks are normally referred to as "streaming networks" [10]. Typical examples include social networks and collaboration networks. Those methods ignore the dynamic nature and are unable to update the vertices' embeddings in accordance with networks' evolution. Second, those methods are *transductive*. They require that all vertices in a graph be present during training in order to generate their embeddings and thus cannot generate embeddings for unseen vertices. In streaming networks that constantly encounter new vertices, the *inductive* capability is essential to support diverse machine learning applications. Third, the time complexity of retraining in those methods usually increases linearly with the number of vertices. This makes simple adaptations of the above methods through retraining computationally expensive, let alone the uncertainty of convergence. Indeed, the few very recent works [11–16] adapted from the above methods require either prior knowledge of new vertices' attributes to be inductive or require retraining on the new graph. This presents a challenge for many high-throughput production machine learning systems that need to generate the

representations of new vertices in real time. In fact, a streaming network's structure may not change substantially within a short period of time, and retraining over the entire graph is usually unnecessary.

To overcome the aforementioned limitations, we propose a novel efficient real-time representation learning framework for streaming graphs. In this framework, a constrained optimization model is formulated to preserve temporal smoothness and structural proximity in streaming representations. We show that the model belongs to quadratic optimization with orthogonal constraints, which in general has no closed-form solution. Therefore, we propose an online approximation algorithm that is able to inductively generate embeddings for newly arrived vertices and has a closed-form expression. In the online algorithm, we divide the task of streaming graph embedding into three sub-tasks: identifying original vertices that are affected most by the new vertices, calculating the embeddings of the new vertices and adjusting the embeddings of the affected original vertices. Since the change of a streaming graph within a short time period, compared with the entire network, is small, the algorithm only updates the representations of a small proportion of vertices. Moreover, such an update does not require to retrain a model or wait for convergence, has low space and time complexity and thus our method is particularly suitable for high-throughput production machine learning systems.

Contributions. Our research contributions are as follows:

(1) We propose a novel online representation learning framework for streaming graphs based on a constrained optimization model. Our model simultaneously takes into consideration temporal smoothness and structural proximity. Our framework is able to calculate representations of unseen vertices without knowing their attributes.

(2) We devise an approximation algorithm that is able to generate representations in real time for vertices arriving in a streaming manner. This algorithm is highly efficient. In particular, it does not require retraining on the entire network or additional rounds of gradient descent. Moreover, we prove that the generated representations are still feasible in the original optimization problem.

(3) We conduct extensive experiments on five real-world data sets to validate the effectiveness and efficiency of our model in both a supervised learning task (i.e., multi-class classification) and an unsupervised learning task (i.e., clustering). The results demonstrate that the proposed framework can achieve comparable or even better performance than those achieved by retraining.

2 RELATED WORK

Static Network Embedding. Recent developments in modeling practical problems in complex systems by machine learning tasks on large graphs have highlighted the need for graph embedding. In graph embedding, each vertex is mapped to a low-dimensional vector, while preserving a graph's structural information. Current studies in this direction can be roughly categorized by different types of structural information preserved in the mapping. LINE [5] and SDNE [6] preserve the first- and second-order proximities, with the difference that SDNE uses highly non-linear functions to model the mapping. Inspired by recent advances in natural language processing, DeepWalk [7] and *node2vec* [2] preserve higher-order proximities by maximizing the conditional probability of observing the contexts of a vertex given its representations. The crucial difference lies in that *node2vec* follows a biased approach to sample contexts. *struct2vec* [8] proposes to preserve the structural identity between nodes in the representation. To achieve this goal, it first creates a new graph based on the structural identity similarity between nodes and then follows a similar method to DeepWalk on the created graph. A very recent method GraphWave [9] makes use of wavelet diffusion patterns by treating the wavelets from the heat wavelet diffusion process as distributions.

Table 1: Comparison between our method and existing ones

Paper	Time complexity	Need retraining	Need attributes
[1, 13]	-	✓	✓
[17]	$O(k^2(\|\mathcal{V}_t\|+k))$	×	✓
[18]	$O\left(\|\mathcal{V}_t\|^2\right)$	×	✓
[19]	$O(\|\mathcal{V}_t\|k^2+k^4)$	✓	×
[14]	$O(\|\mathcal{V}_t\|)$	✓	×
Ours	$O(\beta)$	×	×

Dynamic Network Embedding. Most of the aforementioned studies have focused on static and fixed networks. However, the majority of real-world networks evolve over time and continuously grow. New vertices as well as new edges form in a stream fashion. There are several studies on techniques able to generate embeddings for dynamic graphs. We compare them with our method in terms of three aspects in Table 1: for embedding update at time t (1) the time complexity, (2) whether retraining is needed, and (3) whether vertex attributes is needed. From Table 1 we observe that the time complexity of update at time t is $O(\beta)$, where β is the average degree of graph at time t and can be much smaller than $|\mathcal{V}_t|$, the total number of vertices. Furthermore, the discussion of complexity in [17–19] is assumes a sparse adjacency matrix sparsity, which our methods do not require. Our method has closed-form expression and does not need to re-train the model for embedding updates, and thus does not depend on convergence during training. Finally, our method is able to generate embeddings for new vertices without vertex attributes.

3 PROBLEM FORMULATION

For a streaming graph, we consider new vertices and edges to arrive every with time interval between $t_0 + i\Delta t$ and time $t_0 + (i+1)\Delta t$ for $i \in \{0, 1, \ldots\}$, where t_0 is the initial time and Δt is the interval width. We will use t_i as a shorthand of $t_0 + i\Delta t$. The number of vertices and their edges that arrive within any Δt can be arbitrary. Let $\mathcal{G}_{t_i} = (\mathcal{V}_{t_i}, \mathcal{E}_{t_i})$ denote the graph consisting of vertices $\mathcal{V}_{t_i}$ and edges $\mathcal{E}_{t_i}$ formed before time t_i. Let $\Delta\mathcal{V}_{t_i}$ and $\Delta\mathcal{E}_{t_i}$ be the vertices and their edges formed between time t_i and t_{i+1}. For any time t_i, adding the vertices $\Delta\mathcal{V}_{t_i}$ and the edges $\Delta\mathcal{E}_{t_i}$ to the graph $\mathcal{G}_{t_i}$ leads to the new graph $\mathcal{G}_{t_{i+1}}$ at time t_{i+1}. For example, consider the "Observed" rectangle in Figure 1. Adding the vertices v_4, v_5, v_6 and their edges (depicted by dashed lines) formed between time t_0 and t_1 to $\mathcal{G}_{t_0}$ leads to $\mathcal{G}_{t_1}$. Let $\mathbf{f}_v^{(t_i)} \in \mathbb{R}^k$ be the embedding of vertex $v \in \mathcal{V}_{t_i}$, where the embedding dimension $k \ll |\mathcal{V}_{t_i}|$. Then at any time t_i, the collection of embeddings of vertices arrived before t_i is denoted by $\{\mathbf{f}_v^{(t_i)}\}_{v\in\mathcal{V}_{t_i}}$. Our objective is to generate embeddings

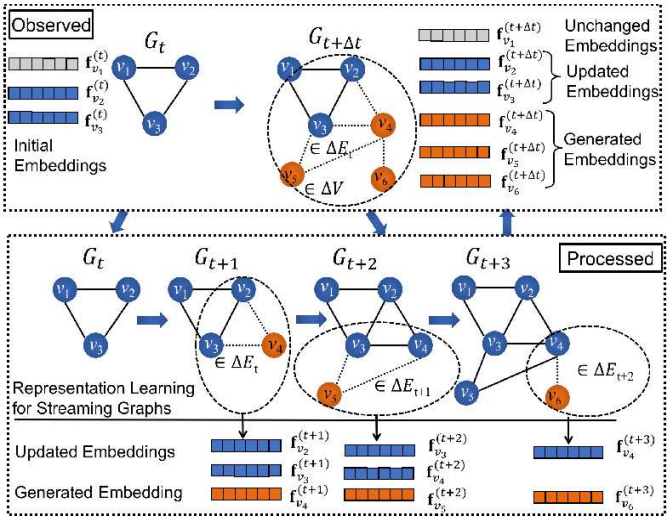


Figure 1: An illustrated example of the proposed embedding technique for streaming graphs with $D = 1$. Gray squares represent unchanged embeddings, blue ones represent updated embeddings for influenced vertices and orange ones represent generated embeddings for new vertices. To be clear, $\mathbf{f}_{v_2}^{(t+\Delta t)} = \mathbf{f}_{v_2}^{(t+1)}$, $\mathbf{f}_{v_3}^{(t+\Delta t)} = \mathbf{f}_{v_3}^{(t+2)}$, $\mathbf{f}_{v_4}^{(t+\Delta t)} = \mathbf{f}_{v_4}^{(t+3)}$, $\mathbf{f}_{v_5}^{(t+\Delta t)} = \mathbf{f}_{v_5}^{(t+2)}$, and $\mathbf{f}_{v_6}^{(t+\Delta t)} = \mathbf{f}_{v_6}^{(t+3)}$.

for the new vertices with a *real-time*, *low-complexity* and *efficient* approach. Now we can formally define the real-time representation learning problem for streaming graphs as follows, using notations described in Table 2.

Definition 1. [Streaming graph embedding] *Consider a graph $\mathcal{G}_{t_0} = (\mathcal{V}_{t_0}, \mathcal{E}_{t_0})$, possibly empty, at initial time t_0. Starting from $i = 0$, a set of vertices $\Delta\mathcal{V}_{t_i}$ and associated edges $\Delta\mathcal{E}_{t_i}$ form in graph $\mathcal{G}_{t_i}$ between time t_i and t_{i+1} and results in a new graph $\mathcal{G}_{t_{i+1}}$ at time t_i. At any time t_{i+1} with $i \in \mathbb{N}$, (1) generate representations $\{\mathbf{f}_v^{(t_{i+1})}\}_{v\in\Delta\mathcal{V}_{t_i}}$ for new vertices $\Delta\mathcal{V}_{t_i}$, and (2) update the representations $\{\mathbf{f}_v^{(t_i)}\}_{v\in\mathcal{V}_{t_i}}$ to $\{\mathbf{f}_v^{(t_{i+1})}\}_{v\in\mathcal{V}_{t_i}}$ for existing vertices $\mathcal{V}_{t_i}$.*

4 METHODS

In this section, we first provide a brief introduction to representation learning for static graphs based on spectral theory and then present our model for streaming graphs. We show that our model belongs to the class of quadratic optimization problem with orthogonality constraints, which has no general closed-form solution. We proposed an approximate solution that has low-complexity, high efficiency and being real-time. The approximated solution is composed of three steps: (1) identify vertices influenced most by the arrival of the new vertices, (2) generate representations of the new vertices, and (3) adjust the representations of the influenced vertices. The approximated solution is inspired by the line-search method on the Stiefel manifold and influence the diffusion process.

4.1 Static Graph Representation Learning

Consider a static graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, .., v_{|\mathcal{V}|}\}$. Each edge in $\mathcal{E}$ is represented by its two ends, i.e., (v_i, v_j). The target of spectral theory based graph representation [20] is to keep the representations of two vertices close if they are connected, a reflection of *graph homophily*. Denote the adjacency matrix of $\mathcal{G}$ by $\mathbf{A}$, where $\mathbf{A}(i,j) = 1$ when $(v_i, v_j) \in \mathcal{E}$ and $\mathbf{A}(i,j) = 0$ otherwise. For graph $\mathcal{G}$, this target can be modelled as the optimization problem below:

$$\min_{\mathbf{F}} \mathcal{L}(\mathbf{F}) = \frac{1}{2}\sum_{i,j=1}^{|\mathcal{V}|} \mathbf{A}(i,j)\left\|\mathbf{f}_{v_i} - \mathbf{f}_{v_j}\right\|^2 \quad s.t.\ \mathbf{F}^\top\mathbf{F} = \mathbf{I}_{k\times k}, \tag{1}$$

where the matrix of embeddings $\mathbf{F} \in \mathbb{R}^{|\mathcal{V}|\times k}$ is:

$$\mathbf{F} = \left[(\mathbf{f}_{v_1})^\top, (\mathbf{f}_{v_2})^\top, \cdots, (\mathbf{f}_{v_{|\mathcal{V}|}})^\top\right]^\top \tag{2}$$

Denote the diagonal matrix of $\mathcal{G}$ by $\mathbf{D}$ and denote its element in the i-th row and j-th column by $\mathbf{D}(i,j)$. Then the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}(i,i) = \sum_{j=1}^{|\mathcal{V}|}\mathbf{A}(i,j)$ and $\mathbf{D}(i,j) = 0$ for $i \neq j$. Belkin et al. [21] show that Eq. (1) can be solved by finding the top-k eigenvectors of the following generalized eigen-problem: $\mathbf{Lx} = \lambda\mathbf{Dx}$. Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{|\mathcal{V}|}$ be the eigenvectors of the corresponding eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_{|\mathcal{V}|}$. It is easy to verify that $\mathbf{1}$ is the only corresponding eigenvector for eigenvalue λ_1. Then the matrix of embeddings can be obtained by $\mathbf{F} = [\mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_{k+1}]$. The time complexity of calculating $\mathbf{F}$ can be as high as $O(k|\mathcal{V}|^2)$ without any sparsity assumption [22].

Notations	Descriptions or Definitions
$\mathcal{G}_{t_i}$	Graph consisting of vertices and edges formed before time t_i
$\mathcal{V}_{t_i}$	Vertices formed before time t_i
$\Delta\mathcal{V}_{t_i}$	Vertices formed between time t_i and t_{i+1}
$\mathcal{E}_{t_i}$	Edges formed before time t_i
$\Delta\mathcal{E}_{t_i}$	Edges formed between time t_i and t_{i+1}
k	Embedding dimension
D	Depth of influence $D = \{1, 2, \dots\}$
$\mathbf{f}_v^{(t_i)}$	$\mathbb{R}^k$ representation of vertex v at time t_i
$\mathbf{A}_{t_i}$	Adjacency matrix of graph $\mathcal{G}_{t_i}$
$\mathbf{D}_{t_i}$	Diagonal matrix of graph $\mathcal{G}_{t_i}$
$\mathbf{L}_{t_i}$	Laplacian matrix of graph $\mathcal{G}_{t_i}$
$\{\lambda_j^{(t_i)}\}_{j=1}^{\lvert\mathcal{V}_{t_i}\rvert}$	Eigenvalues of $\mathbf{D}_{t_i}^{-1}\mathbf{L}_{t_i}$ in ascending order $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{\lvert\mathcal{V}_{t_i}\rvert}$
$\mathbf{x}_j^{(t_i)}$	$\mathbb{R}^{\lvert\mathcal{V}_{t_i}\rvert}$ eigenvector corresponding to λ_j
$\mathcal{I}_{t_i}(m)$	Set of vertices influenced by vertex m in $\mathcal{G}_{t_i}$
$p_{uv}^{(t+1)}$	Probability that v influences u in graph $\mathcal{G}_{t+1}$
$\lvert\cdot\rvert$	Cardinality of a set
$\lVert\cdot\rVert$	The l_2 norm
$\lVert\cdot\rVert_F$	The Frobenius norm
$tr(\cdot)$	Trace of a matrix

Table 2: Notations and Symbols.

4.2 Dynamic Graph Representation Learning

For simplicity of presentation, our explanation for dynamic graph representation learning focuses on the case where only one vertex along with part of its edges is added to a graph each time. In fact, with a solution able to handle a single vertex addition at a time, the addition of multiple vertices can be solved by sequentially processing multiple single vertex additions. This is illustrated by the "Processed" rectangle in Figure 1. Processing the addition of

$\{v_4, v_5, v_6\}$ in a batch can be decomposed into the sequential processing of adding v_4 at $t+1$, v_5 at $t+2$ and v_6 at $t+3$, where $t+i$ simply indicates the virtual order of processing and does not have any practical meaning. Therefore, in below discussion, suppose initially at time $t_0 = 1$, the graph is empty and starting from $t_0 = 1$, there is vertex arrival between t and $t+1$ for any $t \geq t_0$. Also suppose $\Delta t = 1$. Then we denote the single vertex and part of its edges that arrive at time t by v_t and $\Delta\mathcal{E}_t$, respectively. Then we have, $\mathcal{V}_t = \{v_1, v_2, ..., v_{t-1}\}$ and $\mathcal{E}_t = \bigcup_{i=1}^{t-1} \Delta\mathcal{E}_i$.

To solve the problem defined in Definition 1, we propose an optimization problem that needs to be solved at time $t = 2, 3, ...$ The objective function of the optimization problem is designed based on two key properties of the graph streams: *temporal smoothness* and *graph homophily*. First, since only one vertex and its edges arrive per time, the dynamic graph will evolve smoothly, most of the representations of the same vertices at two consecutive time steps should be close. This property is referred to as *temporal smoothness*. This property has also been observed and shown to be helpful to improve representation performance in [13]. Suppose that we are at time $t+1$. Then, this property can be modeled by minimizing the following objective function at any time $t+1$:

$$\mathcal{L}_s^{(t+1)}(\mathbf{F}_{t+1}) := \sum_{v_i \in \mathcal{V}_t} \left\| \mathbf{f}_{v_i}^{(t+1)} - \mathbf{f}_{v_i}^{(t)} \right\|^2, \tag{3}$$

which is the sum of squared ℓ_2-norm representation difference for the same vertices in two consecutive graph snapshots $\mathcal{G}_t$ and $\mathcal{G}_{t+1}$.

Second, the target of representation learning suggests that connected vertices should be embedded to close points in the latent representation space. This property is referred to as *graph homophily*. This property has been reflected in the objective function and constraints of the optimization in Eq. (1). Thus, they should be kept for the new graph $\mathcal{G}_{t+1}$. Formally, this property can be modeled by minimizing the following objective function at time $t+1$:

$$\mathcal{L}_h^{(t+1)}(\mathbf{F}_{t+1}) := \frac{1}{2} \sum_{i,j=1}^{|\mathcal{V}_{t+1}|} \mathbf{A}_{t+1}(i,j) \left\| \mathbf{f}_{v_i}^{(t+1)} - \mathbf{f}_{v_j}^{(t+1)} \right\|^2. \tag{4}$$

To take into account these two properties, we include both $\mathcal{L}_s^{(t+1)}$ and $\mathcal{L}_h^{(t+1)}$ in the final objective function and retain the constraint given in Eq. (1). The optimization problem to solve at time $t+1$ can be summarized as follows.

$$\begin{aligned} &\min_{\mathbf{F}_{t+1}} \mathcal{L}^{(t+1)}(\mathbf{F}_{t+1}) = \gamma_s^{(t+1)} \mathcal{L}_s^{(t+1)}(\mathbf{F}_{t+1}) + \gamma_h^{(t+1)} \mathcal{L}_h^{(t+1)}(\mathbf{F}_{t+1}) \\ &s.t.\ \mathbf{F}_{t+1}^\top \mathbf{F}_{t+1} = \mathbf{I}_{k \times k}, \end{aligned} \tag{5}$$

where, the matrix of embeddings $\mathbf{F}_{t+1} \in \mathbb{R}^{|\mathcal{V}_{t+1}| \times k}$ and $\gamma_s^{(t+1)} = 1/|\mathcal{V}_{t+1}|$ and $\gamma_h^{(t+1)} = 1/(4|\mathcal{E}_{t+1}|)$ are trade-off terms for the temporal smoothness loss functions $\mathcal{L}_s^{(t+1)}$ and graph homophily loss function $\mathcal{L}_h^{(t+1)}$. It is straightforward to observe that $\mathcal{L}_h^{(t+1)}(\mathbf{F}_{t+1})$ is convex in $\mathbf{F}_{t+1}$. Since $\mathcal{L}_s^{(t+1)}(\mathbf{F}_{t+1})$ can be expressed by

$$\begin{aligned} \mathcal{L}_s^{(t+1)}(\mathbf{F}_{t+1}) &= \sum_{v_i \in \mathcal{V}_t} \left\| \mathbf{f}_{v_i}^{(t+1)} - \mathbf{f}_{v_i}^{(t)} \right\|^2 = \left\| \mathbf{J}_{t+1}\mathbf{F}_{t+1} - \mathbf{F}_t \right\|_F^2 \\ &= tr\Big(\big(\mathbf{J}_{t+1}\mathbf{F}_{t+1} - \mathbf{F}_t \big)^\top \big(\mathbf{J}_{t+1}\mathbf{F}_{t+1} - \mathbf{F}_t \big) \Big), \end{aligned} \tag{6}$$

where $\mathbf{J}_{t+1} \in \mathbb{R}^{|\mathcal{V}_t| \times |\mathcal{V}_{t+1}|}$ is:

$$\mathbf{J}_{t+1} := \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}_{|\mathcal{V}_t| \times |\mathcal{V}_{t+1}|}, \tag{7}$$

$\mathcal{L}_s^{(t+1)}(\mathbf{F}_{t+1})$ is also convex. Therefore $\mathcal{L}^{(t+1)}(\mathbf{F}_{t+1})$ the objective function in Eq. (5) is convex in $\mathbf{F}_{t+1}$.

Since the constraints in Eq. (5) are orthogonality constraints, the optimization problem to solve is a general formed quadratic optimization problem under orthogonality constraints. The space defined by the orthogonal constraints is Stiefel manifold. The problem with such format has been widely studied and concluded with no closed-form solution. The state-of-the-art solution is to learn the solution through Riemann gradient approach [23] or line-search method on the Stiefel manifold [24], whose convergence analysis has attracted extensive research attention very recently. However, they are not suitable for streaming setting, because waiting for convergence brings in time uncertainty and gradient-based methods possess unsatisfied time complexity.

4.3 Approximated Algorithm in Graph Streams

Motivated by the aforementioned limitations, we propose an approximated solution that satisfies the low-complexity, efficiency and real-time requirement in a streaming setting. The proposed approximated solution is inspired by an observation of the line-search method. The basic idea of the line-search method for the optimization problem is to search the optimal solution in the tangent space of the Stiefel manifold. We observed that the line-search method based on the polar decomposition-based retraction updates the representation of a vertex through a linear summation of other representations in iterations [24]. In our problem, that means:

$$\mathbf{F}_{t+1}^{(i+1)} = \left(\mathbf{F}_{t+1}^{(i)} + \alpha_i \mathbf{\Gamma}^{(i)} \right) \left[\mathbf{I}_{k \times k} + \alpha_i^2 (\mathbf{\Gamma}^{(i)})^\top \mathbf{\Gamma}^{(i)} \right]^{-1/2}, \tag{8}$$

where α_i is the step size, $\mathbf{\Gamma}^{(i)}$ is the search direction in the tangent space of the Stiefel manifold at iteration i, and $\mathbf{F}_{t+1}^{(i)}$ is the matrix of embedding at iteration i. This inspires us to generate new embedding for a vertex from a linear summation of original embedding for other vertices. Meanwhile, the temporal smoothness in the problem indicates that the embeddings of most vertices would not change a lot. Therefore, to reduce the summation complexity, in the approximated solution, we propose to only update the embeddings of vertices that are influenced by the new vertex. We summarize the steps of the approximated solution as follows: (1) identify vertices influenced most by the arrival of the new vertices, (2) generate embeddings of the new vertex, and (3) adjust the embeddings of the influenced vertices.

The task for the first step can be summarized as: given a vertex, identify the set of vertices that are influenced by it. Similar problems have been widely discussed in the field "influence propagation" and "information diffusion" (see [25] for a survey). A marriage between this field and graph representation learning has been shown very successful in a few recent works [9, 26] for static graphs. Therefore, we apply the Weighted Independent Cascade Model (WICM), one of the most widely used models in this field, to model the influence

spread. Suppose the influence is spread through multiple rounds when vertex v first becomes influenced at round j, it is given a *single* chance to influence a currently uninfluenced neighbor u at round $j+1$. It succeeds with a probability $p_{uv}^{(t+1)}$. The outcome is independent of the history and of v's influence to other vertices. $p_{uv}^{(t+1)}$ is the probability that v influences u in graph $\mathcal{G}_{t+1}$ and can be estimated through

$$p_{uv}^{(t+1)} := \frac{1}{\sum_{i \in \mathcal{V}_{t+1}} \mathbf{A}_{t+1}(i,u)}, \tag{9}$$

where the denominator is the in-degree of vertex u in graph $\mathcal{G}_{t+1}$. If u has multiple already-influenced neighbors other than v, their attempts are sequenced in an arbitrary order. The new influenced vertices will also have a single chance to influence their neighbors in next round. Denote by D total number of rounds determined by us. The set of influence vertices is determined using Algorithm 1.

Algorithm 1: Influenced vertices identification

Input: Graph $\mathcal{G}_{t+1}$, influenced rounds D, new vertex v_t
Output: $\mathcal{I}_{t+1}(v_t)$ set of vertices influenced by new vertex v_t

1 $k = 0, R_k = \{v_t\}$;
2 **while** $k \leq D$ **do**
3 $k = k+1, R_{k+1} = \varnothing, \mathcal{I}_{t+1}(v_t) = \varnothing$;
4 **for** $v \in R_k$ **do**
5 **for** $u \in \mathcal{N}_{t+1}(v)$ **do**
6 Draw $r \sim Bernoulli(p_{uv})$;
7 **if** $r = 1$ **then**
8 $R_{k+1} = R_{k+1} \bigcup \{u\}$;
9 **end**
10 **end**
11 **end**
12 $\mathcal{I}_{t+1}(v_t) = \mathcal{I}_{t+1}(v_t) \bigcup R_{k+1}$;
13 **end**
14 **return** $\mathcal{I}_{t+1}(v_t)$

There are two benefits in computing $\mathcal{I}_{t+1}(v_t)$ using Algorithm 1. First, by adjusting the value of D we can control the time complexity. When $D = 1$, Algorithm 1 will stop after all neighbors of v_t are visited. This is where $O(\beta)$ comes from. Second, the influence from new vertex v_t is not equal among already-arrived vertices. It is reasonable to hope the representation of a vertex influenced less by v_t has a smaller chance to be updated than those influenced more by v_t. This has already been handled by WICM. Compared to vertices close to the v_t, those far-away are less likely to be included in $\mathcal{I}_{t+1}(v_t)$ because to be included, all the outcomes in line 7 of Algorithm 1 must be 1 along an influence spreading path. We also note that $\mathcal{I}_{t+1}(v_t)$ can be stored and thus computed incrementally. All these benefits enable high efficiency in the streaming setting.

After identifying influenced vertices, following the idea inspired by the line-search method, we generate the embedding for a new vertex through a linear summation of the influenced vertices' embeddings and adjust the original embeddings of influenced vertices. This is detailed in Algorithm 2. The quantity α_{t+1} is determined in a way that the orthogonal constraints in Eq. (5) are satisfied:

$$\alpha_{t+1} := 1 - \sqrt{1 - 1/|\mathcal{I}_{t+1}(v_t)|}. \tag{10}$$

Algorithm 2: Representation generation and update

Input: Graph $\mathcal{G}_t$, newly arrived vertex v_t, newly arrived edges $\Delta\mathcal{E}_t$, matrix of embeddings $\mathbf{F}_t$
Output: Updated matrix of embeddings $\mathbf{F}_{t+1}$

Update graph: $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t \cup \{v_t\}$ and $\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t \cup \Delta\mathcal{E}_t$;
Calculate representation for new vertex v_t by:
$\mathbf{f}_{v_t}^{(t+1)} = \frac{1}{|\mathcal{I}_{t+1}(v_t)|} \sum_{u \in \mathcal{I}_{t+1}(v_t)} \mathbf{f}_u^{(t)}$;
Adjust representations for already-arrived vertices:
for $u \in \mathcal{V}_t$ **do**
  $\mathbf{f}_u^{(t+1)} = \begin{cases} \mathbf{f}_u^{(t)} - \alpha_{t+1}\mathbf{f}_{v_t}^{(t+1)} & u \in \mathcal{I}_{t+1}(v_t) \\ \mathbf{f}_u^{(t)} & o.w. \end{cases}$
end
return $\mathbf{F}_{t+1}$

Algorithm 2 ensures that the embedding of a vertex is generated when it arrives and will be updated when it is influenced by some vertices that come after it. That makes a connection between vertices that arrive at different orders and preserves the temporal pattern in later update. Since the algorithm will only update the embeddings of influenced vertices, different from those solutions that suffer time uncertainty from retraining, the proposed algorithm guarantees to output $\mathbf{F}_{t+1}$ after $|\mathcal{I}_{t+1}(v_t)|$ operations. Therefore, the time complexity of Algorithm 2 is $O\left(|\mathcal{I}_{t+1}(v_t)|\right)$ and is expected to have small variance in running time. The value of $|\mathcal{I}_{t+1}(v_t)|$ can be controlled through changing value of D. Thus this algorithm enables trade-off between complexity and performance in the streaming setting. As discussed in Section 2, it can be as low as $O(\beta)$ with β denoting the average degree of the graph and $D = 1$.

5 EXPERIMENTS

In this section, we conduct experiments of both multi-class vertex classification and network clustering on five data sets to evaluate the effectiveness and efficiency of the proposed method. We use the indices of vertices as their arriving order and generate their arrived edges at each time randomly to model the streaming scenario. We evaluate our method in terms of the performance of the learning tasks and the running time to generate vertex embeddings. The experiments are structured to answer the following questions:

- Effectiveness: compared to state-of-the-art retraining based approaches, how does the proposed approach perform in supervised learning and unsupervised learning in the streaming setting?
- Efficiency: compared to state-of-the-art retraining based approaches, how fast is the proposed solution able to generate new embeddings?
- Scalability and stability: how stable and scalable is the proposed solution in different-scale networks?

5.1 Data Sets

We use the following five real data sets to validate the propose framework. All of them are publicly available and have been widely

used in previous research of static and dynamic graph embedding. The statistics of the datasets are summarized in Table 3.

- **Blog** was collected from the BlogCatalog website, which manages bloggers and their posted blogs. Bloggers follow each other to form network edges. Bloggers categorize their blogs under predefined classes, which are taken as the ground truth of class labels.
- **CiteSeer** is a literature citation network for the selected papers indexed in CiteSeer. Papers are considered as vertices. The paper citation relations are considered as the links in the network and papers are classified into the following six classes: Agents, Artificial Intelligence, Database, Information Retrieval, Machine Learning and Human-Computer Interaction.
- **Cora** also represents a citation network, whose vertices represent publications from 7 classes: Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, and Theory. Each link is a citation relationship between the two publications.
- **Flickr** was collected from Flickr, an image sharing website hosting images uploaded by users. Users in Flickr interact with others to form edges. User can subscribe different interest groups, which correspond to the class labels. The interest groups, for instance, are “black and white photos".
- **Wiki** contains 2,405 documents from 17 classes and 17,981 links between them.

Table 3: Dataset statistics

Dataset	# of vertices	# of edges	# of classes
Blog	5,196	171,743	6
CiteSeer	3,312	4,732	6
Cora	2,708	5,429	7
Flickr	7,575	239,738	9
Wiki	2,405	17,981	17

5.2 Comparison with Baseline Methods

We compare our approach with the following four graph embedding algorithms. Since they are designed for static graph embedding, the retraining based utility of them has achieved similar performance in graph embedding tasks for dynamic graphs compared to dynamic methods. Many studies on dynamic graph embedding have used them as baseline methods. For instance, *node2vec* in [19, 27–29], NetMF in [17, 19], DeepWalk in [17, 19, 29]. Except for those already tested in existing works, we also compare our solution with a new framework *struct2vec*. For each baseline, a combination of their hyper parameters are tested and the one achieving the best performance is reported as their performance. To be fair, our solution use the same values for the shared hyper parameters. In the following, we refer to “walk length" as wl, “window size" as ws, and representation dimensions as d. The values of hyper parameters for baselines are obtained through grid search of different combinations: $d \in \{10, 20, ..., 200\}$, $wl \in \{10, 20, 30, 40\}$, $ws \in \{3, 5, 7, 10\}$. where the finally chosen values are $d = 90$, $wl = 10$, $ws = 7$.

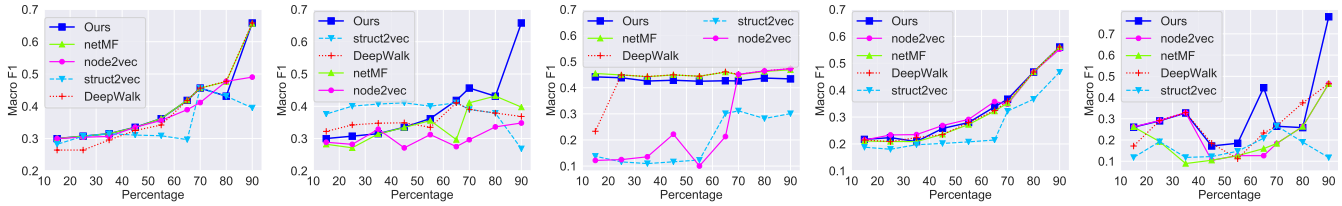
- **NetMF** [30] obtains graph embeddings through explicitly factorizing the closed-form matrices It has been shown to outperform LINE [5] on several benchmark data sets.
- **DeepWalk** [7] learns graph embeddings by preserving higher-order structural proximity between vertices in the latent space. A pair of vertices are considered similar if they are close in truncated random walks.
- **node2vec** is equipped with biased random walk to provide a trade-off between BFS and DFS. Compared to DeepWalk, it has a more flexible strategy to explore neighborhoods. The ranges of its unique hyper parameters are experimented with: $p \in \{0.5, 1, 1.5, 2\}$, and $q \in \{0.5, 1, 1.5, 2\}$, where $p = 1$, $q = 1$ and the number of walks is 10 are finally reported.
- **struct2vec** [8] learns embeddings by preserving the structural similarities between vertices in the embedded space, where a hierarchical multi-layer graph is used to encode vertex structural similarities at different scales.

5.3 Supervised Tasks - Vertex Classification

To evaluate the effectiveness of the proposed model, we first compare the performance of our method with different baseline methods on the vertex classification task. The vertex embeddings are fed into a one-vs-rest logistic regression classifier with L2 regularization. In our context, the training percentage (say 20%) specifies an initial portion of arriving vertices for which the offline spectral theory based method in (1) to generate the initial embeddings and then follow Algorithm 2 to learn embeddings for vertices arriving thereafter. For a vertex that arrives in the testing phase, only its embedding obtained upon arrival is used in testing, although later this may be updated, e.g., when there is arrival of its neighbors. As comparison, baseline methods are retrained on all already arrived vertices to generate embeddings for new vertices. Correspondingly, the classifier for the our method will be trained on the embeddings for the first 20% arrived vertices and tested on the remaining 80% arrived vertices. In other words, the classifiers, the baseline methods and the proposed method must be trained and tested on same percentage split of arrival vertices.

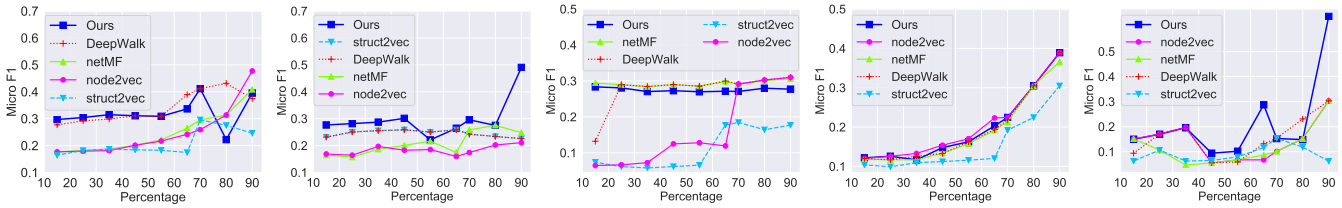
5.3.1 Discussion of Effectiveness. We use Micro-F_1 and Macro-F_1 as the evaluation metrics. Figure 2 and Figure 3 compare the Macro-F_1 and Micro-F_1 performance in testing, respectively, with varying percentages of data used for training. We observe that overall, the performance of both the proposed method and the baseline methods improves as the percentage of training increases. This is because as the larger percentage of data is used for training, the embeddings obtained by the proposed methods and retraining based baseline methods are more similar to those by static methods. It can be also observed that our solution achieves almost the same or even slightly better Micro-$F1$ and Macro-F_1 under varying percentages. For example, on CiteSeer, Cora and Wiki data set, our method slightly outperforms baseline methods for most percentages. On Blog and Flickr data set, our method achieves almost the same Micro-F_1 and Macro-F_1 for most percentages.

The observation that retraining based baseline methods sometimes are slightly worse than the proposed method indicates that when a graph is highly incomplete, embeddings generated by retraining based baseline methods that considered global structural information may not be as reliable as the proposed method that only considers local updates. This is because the global information baseline methods are highly biased particularly when only a



(a) Blog (b) CiteSeer (c) Cora (d) Flickr (e) Wiki

Figure 2: Comparison of vertex multi-class classification performance in Macro-F_1 with $D = 1$.



(a) Blog (b) CiteSeer (c) Cora (d) Flickr (e) Wiki

Figure 3: Comparison of vertex multi-class classification performance in Micro-F_1 with $D = 1$.

small proportion of the whole graph has arrived. This effect slightly degrades the baseline performance in some cases.

5.3.2 Discussion of Efficiency. We empirically evaluate the running times of different methods in Figure 4. Note that the y-axis is in log scale. The running times for the proposed method count both the times to generate training vertices and testing vertices. Since the variation of training percentages influence the running times, we collect running times over all training percentages. The running times for the baseline methods count the times to generate embeddings for all vertices. We observe that, in general when same number of embeddings are generated, the running time of our solution is much smaller. If we compare them under generation at each t, the running time of our solution will be further shorter because our method has lower time complexity as discussed in Section 2. Meanwhile, along with the statistics in Table 3, we note the increased running times of the proposed method is not as large as other baseline methods when the network size increases. That empirically demonstrates scalability of the proposed method. This is because that all retraining baseline methods need to wait for convergence of retraining, whose uncertainty increases as the network size increases, while our method only need to update over its neighbors and guarantee to stop after certain steps.

5.4 Unsupervised Tasks - Network Clustering

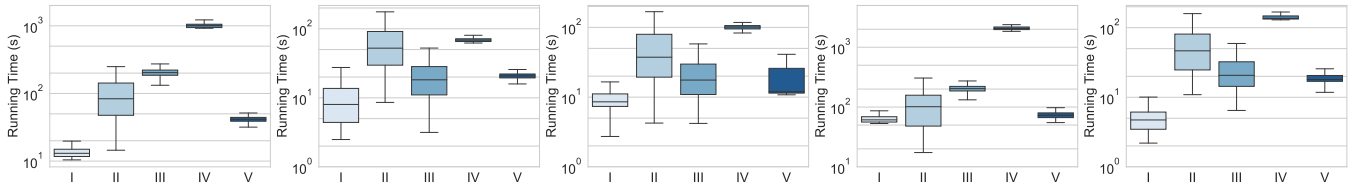
Next, we assess the effectiveness of different vertex representations on an unsupervised learning task - network clustering. Since the variation of training percentages influence the performance, we compare the average clustering performance over all training percentages. We use the same embeddings used in vertex classification task. Thus our method's running time is also illustrated in Figure 4. We perform K-means clustering based on the embeddings generated by our method and different baselines. K is chosen to equal the number of classes in each data set. K-means algorithm is repeated 10 times and the average results are reported since K-means may converge to the local minima due to different initialization. We use normalized mutual information (NMI) and completeness score as the performance metrics. They help quantify how close the clustering results are to the ground-truth class belongings. The computation of the two evaluation metrics can be expressed below:

$$NMI := \frac{2I(C;K)}{H(C)+H(K)}, \quad Completeness := 1 - \frac{H(K|C)}{H(K)},$$

where C denotes the class assignment, K denotes the clustering assignment, $I(X;Y) = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y) log p(x,y)/p(x)p(y)$ is the mutual information between random variable X and Y and $H(X) = -\sum_{x\in\mathcal{X}} p(x) log p(x)$ is the entropy. The results are summarized in Table 4. Again it can be seen that our method achieves comparable or slightly better performance. For example, our method achieves slightly better performance on Cora and Flickr. As a reminder, Flickr network is the largest among all. Please refer Section 5.3.1 for discussion of the reasons.

6 CONCLUDING REMARKS

We proposed an efficient online representation learning framework for graph streams, in which new vertices and edges arrive as a stream. The framework is inspired by incrementally approximating the solution to a constructed constrained optimization problem, which preserves temporal smoothness and structural proximity in resultant representations. Our approximating solution has closed form, high efficiency, and low complexity, and remains feasible under orthogonality constraints. To validate the effectiveness of our model and learning algorithm, we conducted experiments on five real-world networks for both supervised and unsupervised learning tasks (multi-class classification and clustering) with four baseline methods. Experimental results demonstrate that compared



(a) Blog (b) CiteSeer (c) Cora (d) Flickr (e) Wiki

Figure 4: Comparison of running time in seconds (I-Ours, II-DeepWalk, III-Node2Vec, IV-Struct2Vec, V-NetMF).

Table 4: Comparison of performance on clustering %

	Blog		CiteSeer		Wiki		Cora		Flickr	
	Completeness	NMI	Completeness	NMI	Completeness	NMI	Completeness	NMI	Completeness	NMI
Ours	16.48	**26.71**	16.48	22.46	16.48	26.71	**34.50**	**34.62**	**16.53**	**20.44**
DeepWalk	**17.77**	20.01	**17.77**	20.01	11.78	11.72	34.62	34.30	16.46	17.31
netMF	0.55	0.67	0.25	0.27	6.85	7.17	6.99	7.49	1.91	2.05
node2vec	16.78	22.93	17.55	**22.93**	**16.60**	**27.77**	31.77	31.60	15.82	21.65
struct2vec	4.35	6.44	2.34	2.44	3.57	4.74	9.80	7.94	7.23	8.04

with several state-of-the-art techniques, our approach achieves comparable performance to that of retraining the entire graph with substantially less running time.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant No. IIS-1848596

REFERENCES

[1] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 2017.
[2] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.
[3] Ye Li, Chaofeng Sha, Xin Huang, and Yanchun Zhang. Community detection in attributed graphs: An embedding approach. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 338–345, 2018.
[4] Zhu Cao, Linlin Wang, and Gerard de Melo. Link prediction via subgraph embedding-based convex matrix completion. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 2803–2810, 2018.
[5] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, 2015.
[6] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234, 2016.
[7] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.
[8] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 385–394, 2017.
[9] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. Learning structural node embeddings via diffusion wavelets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1320–1329, 2018.
[10] Shiyu Chang, Yang Zhang, Jiliang Tang, Dawei Yin, Yi Chang, Mark A Hasegawa-Johnson, and Thomas S Huang. Positive-unlabeled learning in streaming networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 755–764, 2016.
[11] Lun Du, Yun Wang, Guojie Song, Zhicong Lu, and Junshan Wang. Dynamic network embedding: An extended approach for skip-gram based network embedding. In *Proceedings of the 17th International Joint Conferences on Artificial Intelligence*, pages 2086–2092, 2018.
[12] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.
[13] Xi Liu, Muhe Xie, Xidao Wen, Rui Chen, Yong Ge, Nick Duffield, and Na Wang. A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games. In *IEEE International Conference on Data Mining (ICDM)*, 2018.
[14] Jianxin Ma, Peng Cui, and Wenwu Zhu. Depthlgp: Learning embeddings of out-of-sample nodes in dynamic networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 370–377, 2018.
[15] Giang Hoang Nguyen, John Boaz Lee, Ryan A Rossi, Nesreen K Ahmed, Eunyee Koh, and Sungchul Kim. Continuous-time dynamic network embeddings. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 969–976. International World Wide Web Conferences Steering Committee, 2018.
[16] Qixiang Wang, Shanfeng Wang, Maoguo Gong, and Yue Wu. Feature hashing for network representation learning. In *IJCAI*, pages 2812–2818, 2018.
[17] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. Attributed network embedding for learning in a dynamic environment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 387–396, 2017.
[18] Ling Jian, Jundong Li, and Huan Liu. Toward online node classification on streaming networks. *Data Mining and Knowledge Discovery*, 32(1):231–257, 2018.
[19] Dingyuan Zhu, Peng Cui, Ziwei Zhang, Jian Pei, and Wenwu Zhu. High-order proximity preserved embedding for dynamic networks. *IEEE Transactions on Knowledge and Data Engineering*, 2018.
[20] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018.
[21] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591, 2002.
[22] Johann Paratte and Lionel Martin. Fast eigenspace approximation using random signals. *EPFL-ARTICLE*, 2017.
[23] Zhiqiang Xu and Xin Gao. On truly block eigensolvers via riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 168–177, 2018.
[24] Huikang Liu, Weijie Wu, and Anthony Man-Cho So. Quadratic optimization with orthogonality constraints: explicit łojasiewicz exponent and linear convergence of line-search methods. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 1158–1167, 2016.
[25] Sancheng Peng, Yongmei Zhou, Lihong Cao, Shui Yu, Jianwei Niu, and Weijia Jia. Influence analysis in social networks: a survey. *Journal of Network and Computer Applications*, pages 17–32, 2018.
[26] Yuan Zhang, Tianshu Lyu, and Yan Zhang. Cosine: Community-preserving social network embedding from information diffusion cascades. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 2620–2627, 2018.
[27] Franco Manessi, Alessandro Rozza, and Mario Manzo. Dynamic graph convolutional networks. *arXiv preprint arXiv:1704.06199*, 2017.

[28] Rakshit Trivedi, Mehrdad Farajtbar, Prasenjeet Biswal, and Hongyuan Zha. Representation learning over dynamic graphs. *arXiv preprint arXiv:1803.04051*, 2018.

[29] Le-kui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. Dynamic network embedding by modeling triadic closure process. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 571–578, 2018.

[30] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 459–467, 2018.

Integrating Bayesian and Neural Networks for Discourse Coherence

Jinhua Peng, Zongyang Ma, Di Jiang, Hua Wu
Baidu Inc. China
{pengjinhua,mazongyang,jiangdi,wu_hua}@baidu.com

ABSTRACT

In dialogue systems, discourse coherence is an important concept that measures semantic relevance between an utterance and its context. It plays a critical role in determining the inappropriate reply of dialogue systems with regard to a given dialogue context. In this paper, we present a novel framework for evaluating discourse coherence by seamlessly integrating Bayesian and neural networks. The Bayesian network corresponds to Coherence-Pivoted Latent Dirichlet Allocation (cpLDA). cpLDA concentrates on generating the fine-grained topics from dialogue data and takes both local and global semantics into account. The neural network corresponds to Multi-Hierarchical Coherence Network (MHCN). Coupled with cpLDA, MHCN quantifies discourse coherence between an utterance and its context by comprehensively utilizing original texts, topic distribution and topic embedding. Extensive experiments show that the proposed framework yields superior performance comparing with the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

Bayesian Network, Multi-Hierarchical Coherence Network, Global Semantics, Dialogue systems, Discourse Coherence

ACM Reference Format:
Jinhua Peng, Zongyang Ma, Di Jiang, Hua Wu. 2019. Integrating Bayesian and Neural Networks for Discourse Coherence. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316582

1 INTRODUCTION

Discourse coherence is a concept that measures semantic relevance between a natural language utterance and its context [14]. In the dialogue system, it usually retrieves or generates a set of responses based on a user's utterance. The function of discourse coherence is to rank and select the most appropriate one from the above candidate responses as the reply. This is crucial for the dialogue system, since inconsistent response might hurt the user's experience when interacting with a chatbot. In Figure 1, we present two cases to illustrate the utility of discourse coherence. In each case, the last utterance is a candidate reply of the dialogue system and all its previous utterances are the context. In the first case, *You almost sleep for 11 hours everyday!* is consistent with the theme of the context. On the contrary, in the second case, *Gigantic appetite* seems to be a dramatic shift of the dialogue theme and is not suitable for the given context. By qualitatively evaluating the semantic relevance between candidate replies and the context, discourse coherence enables dialogue systems to filter out inappropriate replies and pinpoint the appropriate ones.

		Chinese	English
Coherent Case	Context	A: … B: … A: 别睡了,快点起来开工了 B: 俺好累好困啊	A: … B: … A: Don't sleep, get up to work! B: I am so tired and sleepy.
		A: 每天睡十一小时	A: You almost sleep for 11 hours everyday!
Incoherent Case	Context	A: … B: … A: 甜点好吃,使我无法自拔 B: 这个看着就好吃啊	A: … B: … A: The dessert tastes good, and I like it very much! B: It looks very delicious!
		A: 超级惊人的饭量	A: Gigantic appetite.

Figure 1: Real-life Dialogue Cases. The original Chinese texts are translated into English to enhance readability.

However, to the best of our knowledge, little work has been done to design specialized models for discourse coherence. Existing work [14] usually utilizes primitive textual information and conventional discriminative/generative models for evaluating discourse coherence, while ignores the topical interaction between utterances. Since these models are not initially proposed for discourse coherence, the capabilities of these methods are heavily limited in real-life applications. Considering the complicated semantic nature of dialogue data, we propose a novel framework that takes full advantage of Bayesian and neural networks for discourse coherence. The framework consists of two major components: the Bayesian component and the neural component, which are described as follows.

Table 1: Notations for cpLDA

z_i	the topic of sentence i
$\mathbf{z}_{-i}$	topics of all the sentences except that of sentence i
$\mathbf{v}_i$	set of words in sentence i
n_{dk}	number of words in document d assigned with topic k
n_{kv}	number of word v assigned with topic k
n_{di}	number of words in sentence i of document d
n_{iv}	number of word v in sentence i

The Bayesian component, which is named as cpLDA (short for Coherence-Pivoted Latent Dirichlet Allocation), encodes our prior

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316582

knowledge of dialogue data into topic-level semantics. cpLDA ensures both global and local semantic consistency for latent topic discovery. Hence, the resultant topics of cpLDA are influenced by the global word relations as well as the local structures within dialogue data (i.e., words in a sentence are constrained to one unique topic). cpLDA is well designed for mining latent topics from dialogue data, in which each utterance is relatively short and lacks of information. On the other hand, the neural component is a deep neural network model, naming MHCN (short for Multi-Hierarchical Coherence Network), equipped with the generated topics from cpLDA for distinguishing the coherent utterances from incoherent ones given the context. MHCN is essentially a three-layer Neural Network (i.e., word-level, sentence-level and combination-level layers), and takes various factors into account, including the Bag-of-Words representation and topic distribution of each utterance.

In order to verify the effectiveness of the proposed framework, we quantitatively compare it with the state-of-the-arts in terms of a variety of metrics. Extensive experimental results show that the proposed framework achieves superior performance and applicability in the dialogue system for the task of response selection.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we discuss the technical details of our framework consisting of cpLDA and MHCN models. In Section 4, we report the experimental results of our proposed models for discourse coherence. Finally, we conclude the paper in Section 5.

2 RELATED WORK

In this section, we begin with a brief overview of the studies on discourse coherence. We then survey the related work about topic models.

The study of discourse coherence is motivated by many work about linguistic theories. Guinaudeau and Strub ([10]) extended the entity-grid model by converting the matrix to the bipartite graph, which is able to address computational complexity and data sparsity problems occurred in the entity-grid model. In [9], the authors combined the entity-grid model and the HMM-based content model in a unified framework and demonstrated that their proposed model outperforms entity-grid model on the task of sentence ordering(2007). Recently, Neural Net Models are proposed in [14] for the task of discourse coherence. The authors introduced two classes of neural models: discriminative models for distinguishing coherent discourse from incoherent one and generative models for producing coherent discourse. More specifically, the generative models are derived from the variational latent variable models, which treat the topic z of current utterance as a stochastic latent variable conditioned on all previous utterances. With the topic latent variable, these generative models perform best on the task of sentences reordering. Dissimilar to the study in [14], we explicitly incorporate the topical embedding into our MHCN model. That is, the topics are preliminarily discovered by cpLDA, which could be more flexible to add into the MHCN model.

The proposed cpLDA can be considered as a deeply restructured variant of LDA. LDA was firstly introduced in [4], where it was developed for discovering hidden topics from a large collection of documents. Various extensions are proposed due to the promising performance of LDA. In the following, we name a few. Dynamic Topic Model [3] incorporates the time attribute, assuming that topics are evolving with time. That is, the same topic can be represented by different sets of keywords during different time periods. Supervised Topic Model was proposed in [8], where topics are regularized by the response type (i.e., document label). In LDA, topics are regarded as independent factors. However, in common sense, topics are usually related with each other. To address this issue, Correlated Topic Model [2] models the relations between topics in a more natural way. Instead of assigning a topic to a word, in sentenceLDA [11], a unique topic is assigned to all the words in a sentence.

3 A FRAMEWORK INTEGRATING BAYESIAN AND NEURAL COMPONENTS

In this section, we present our framework, which consists of a Bayesian component (cpLDA) and a neural component (MHCN).

3.1 cpLDA

Here we discuss the technical details of cpLDA. In Section 3.1.1, we discuss how to impose the local semantics consistency by modeling the "sentences" in the generative process of cpLDA. In Section 3.1.2, we discuss how to achieve global semantics consistency by utilizing the knowledge about word relations.

3.1.1 Local Semantics Consistency. cpLDA associates a latent topic with each observation, which is the occurrence of a "sentence" and its "words" in a particular "document". We utilize d to denote a "document", w a "word", s a "sentence" and z a "latent topic". θ_d and ϕ_k represents the topic distribution of document d and the word distribution of topic k respectively. We also introduce the hyper-parameters α and β. Based on these notations, we present the generative process of cpLDA as follows:

- for each topic z_k, generate the word distribution $\phi_k \sim Dirichlet(\beta)$;
- for each document d, generate the topic distribution $\theta_d \sim Dirichlet(\alpha)$;
- for each sentence s in d:
  - draw a topic $z_k \sim Multinomial(\theta_d)$;
  - draw each word $w \sim Multinomial(\phi_k)$;

In the generative process above, we constrain that the words of a sentence are generated by the same topic, in order to ensure the local semantics consistency. Hence, cpLDA utilizes "sentences" rather than "words" as the basic units for topic assignment. For inferring cpLDA, we need to compute the posterior distribution of the hidden variables z, θ and ϕ as follows:

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

However, due to the coupling between θ and ϕ, this distribution is intractable to compute. Gibbs Sampling is an alternative algorithm for estimating the latent variables in cpLDA. The condition probability $p(z_i = k|\mathbf{z}_{-i}, \mathbf{w})$ is shown in Table 2.
The notations are explained in Table 1. Though Gibbs sampler is a simple method to infer cpLDA, it suffers from the high computational complexity, which limits its usage on large-scale data

Table 2: Conditional probability of topic z_i

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) = \frac{n_{dk} + \alpha_k}{\sum_{k'=1}^{K}(n_{dk'} + \alpha_{k'})} \frac{\Gamma(\sum_{v'=1}^{V}(n_{kv'} + \beta_{v'}))}{\Gamma(\sum_{v'=1}^{V}(n_{kv'} + \beta_{v'}) + n_{di})} \prod_{v' \in \mathbf{v}_i} \frac{\Gamma(n_{kv'} + \beta_{v'} + n_{iv'})}{\Gamma(n_{kv'} + \beta_{v'})}$$

$$p'(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) = \frac{n_{dk} + \alpha_k}{\sum_{k'=1}^{K}(n_{dk'} + \alpha_{k'})} \prod_{v \in \mathbf{v}_i} \frac{n_{kv} + \beta_v}{\sum_{v'=1}^{V}(n_{kv'} + \beta_{v'})}$$

set. In this paper, we introduce Metropolis-Hastings (MH) algorithm for cpLDA. Compared with Gibbs Sampling, when sampling a topic z, the time complexity of Gibbs Sampling is $O(K)$ (K represents the number of topics), while that of MH algorithm is approximately reduced to $O(1)$. The approximate conditional probability $p'(z_i = k|\mathbf{z}_{-i}, \mathbf{w})$ is shown in Table 2.

In order to proceed MH algorithm, several proposals are required. We define two proposals: Doc-Topic Proposal and Topic-Term Proposal. Doc-Topic Proposal is formulated as $\rho_d(k) \propto \frac{n_{dk}+\alpha_k}{\sum_{k'=1}^{K}(n_{dk'}+\alpha_{k'})}$ and the acceptance probability of topic transition from k to k' is $min\{1, \frac{p(k')\rho_d(k)}{p(k)\rho_d(k')}\}$. $p(k)$ can be selected according to the first conditional probability in Table 2. Similarly, the Topic-Term Proposal is formulated as $\rho_v(k) \propto \frac{n_{kv}+\beta_v}{\sum_{v'=1}^{V}(n_{kv'}+\beta_{v'})}$. We further create the alias table to accelerate the sampling process. When using the alias table, generating a topic z only requires two uniform samples, more details of alias table can be referred to [18].

3.1.2 Global Semantics Consistency. The previous section illustrates how to model the "sentences" in cpLDA and details the approach of inferring the parameters via the MH algorithm. In this subsection, we discuss how to utilize global semantics consistency in cpLDA. The knowledge about word relations can be obtained from many sources such as human-engineering ontology [16] and automatically built knowledge bases [7]. In the scenario of dialogue system, high-quality word relations can be easily obtained by computing the similarity of word embedding vectors from an external knowledge base (e.g., microblog).

Assuming that the vocabulary size is W, the word relations are characterized by a $W \times W$ matrix $\mathbf{R}$. We denote the weight of the relation between two words w_a and w_b as R_{ab} and calculate it as follows:

$$R_{ab} = sim(v_a, v_b).$$

where v_a and v_b denote the word embedding vectors of w_a and w_b respectively.

We proceed to discuss the strategy of utilizing $\mathbf{R}$ in cpLDA. We want the probability $p(w|z_k)$ (i.e., ϕ_k^w) to be dependent on the word relation information stored in $\mathbf{R}$. Here we use a quadratic-form influence term with a trade-off factor τ. Formally, given $\mathbf{R}$, we adjust the topic-word distribution $p(w|z_k)$ as follows:

$$p(w|z_k) \leftarrow p(w|z_k) + \tau \frac{p(w|z_k) \sum_{i=1}^{W} R_{iw} p(i|z_k)}{P(\cdot|z_k)^T \mathbf{R} P(\cdot|z_k)}. \quad (1)$$

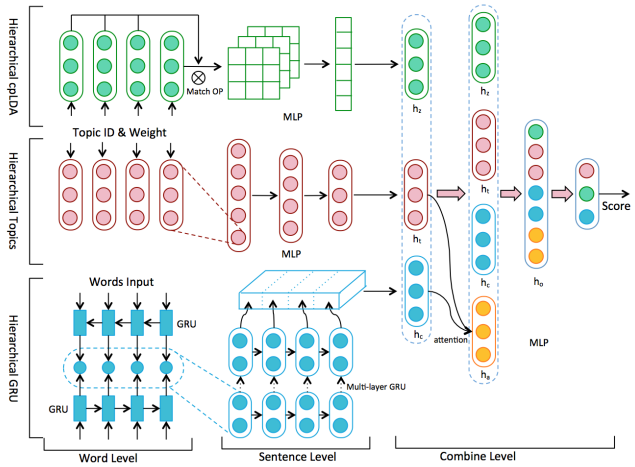


Figure 2: Multi-Hierarchical Coherence Network

It is easy to see that the adjusted $p(w|z_k)$ is influenced by the prominence of the other "words" that has relations with w in $\mathbf{R}$. In practice, Eq. (1) is applied after MH iteration until convergence is achieved. Since we are only interested in relatively frequent words from the vocabulary, $\mathbf{R}$ will be a sparse matrix and hence computations are efficient for $\mathbf{R}$ to be used in practice.

In summary, the global and local semantics consistency modeled in cpLDA fits the characteristics of dialogue data well. The utterance in the dialogue is one type of User Generated Content (UGC) data and each multiple-turn conversation could be treated as the "virtual document". Compared with the traditional web and news documents, this "virtual document" is much shorter with a few sentences, indicating that it contains relatively poor information. Therefore, in cpLDA, globally modeling word relations (global semantics coherence) is able to introduce the external knowledge and enrich the representation of the utterances. On the other hand, utterances are informal text, and full of abbreviations and slangs. That is, constraining each utterance with one unique topic (local semantics coherence) is beneficial for focusing on the key semantics of the utterance.

3.2 MHCN: A Coherence Network Model with Topic Embedding and Topic Attention

In this section, we describe our MHCN in details. MHCN is equipped with the GRU cells, and at the same time, it accepts the input of

topic embeddings generated from the cpLDA model. Let $\mathbf{c}$ denote a sequence of coherent utterances in one session of the dialogue system. More specifically, $\mathbf{c} = \{u_1, u_2, ...u_L\}$, where u and L represents the utterance and the number of utterances respectively. u consists of a sequence of words $\{w_{u,1}, w_{u,2}, ...w_{u,n}\}$. Each word w is described by an embedding vector with the fixed dimensionality. Since the aim of this research is to calculate the coherence degree between the current utterance and all its previous contextual utterances in the same discourse, to make clear, we also use r (response) to denote u_L and $\mathbf{c}^{-r}$ to denote the contextual utterances. The data set is represented as the collection of triple $(\mathbf{c}_i^{-r}, r_i, y_i)$, where y_i is the label. $y_i = 1$ indicates that the response r_i is coherent with the contextual utterances $\mathbf{c}_i^{-r}$. Given the $(\mathbf{c}_i^{-r}, r_i)$ pair, MHCN model $F(\mathbf{c}_i^{-r}, r_i)$ attempts to calculate the coherence score y_i'.

The architecture of MHCN model is illustrated in Figure 2. MHCN is built with hierarchical layers (i.e., word-level, sentence-level and combination-level layers). MHCN accepts three types of input factors: topic ID, topic distribution of the utterance and word sequence of the utterance. (1) For the topic distribution, the similarity scores of the topic distributions between the response r and each utterance in $\mathbf{c}^{-r}$ are calculated. (2) The topic ID is inferred from the cpLDA. Observe that though cpLDA assumes that one utterance only occupies one topic, we concatenate the multiple sampling results for each utterance. In that case, each utterance has multiple topic IDs. The topic ID is treated as the "topical word" in our research and fed into the model as topical word embedding. (3) Each utterance is represented by a sequence of word embeddings, which are conventional embeddings and not pre-trained. Through the GRU cell (the word-level layer), the hidden representation of each utterance u is obtained. Then, these hidden utterance representations are propagated to another GRU cell (sentence-level layer). After this encoding process, the model yields the hidden representation of $\mathbf{c}$. Note that both of the GRU cells (i.e., word-level and sentence-level layers) are bidirectional. The third layer is the combination-level layer, we perform a attention mechanism to obtain utterance attention weight distributions over the topic. In this layer, all these three factor hidden representations and the topic attention weight are combined to calculate the last score.

In MHCN, we could obtain the hidden vector $h_{u,i}$ of utterance u_i by feeding a sequence of embedding word vectors $\{v_{u,1}, v_{u,2}, ...v_{u,n}\}$ into the GRU cell. $h_{u,i}$ can be calculated as follows:

$$
\begin{aligned}
a_i &= \sigma(\mathbf{W_a} v_{u,i} + \mathbf{U_a} h_{u,i-1}) \\
b_i &= \sigma(\mathbf{W_b} v_{u,i} + \mathbf{U_b} h_{u,i-1}) \\
h'_{u,i} &= tanh(\mathbf{W_h} v_{u,i} + \mathbf{U_a}(b_i \odot h_{u,i-1})) \\
h_{u,i} &= a_i \odot h'_{u,i} + (1 - a_i) \odot h_{u,i-1}
\end{aligned}
$$

where $\mathbf{W_a}, \mathbf{U_a}, \mathbf{W_b}, \mathbf{U_b}, \mathbf{W_h}$ and $\mathbf{U_a}$ are the parameters to be learned. a_i and b_i denote the update gate and reset gate respectively.

Similarly, the hidden vector h_c of $\mathbf{c}$ is obtained by feeding a sequence of hidden vector of utterances $h_{u_1}, h_{u_2}, ...h_{u_n}$. The advantage of using GRU cells is that it enables to keep and propagate key information as well as filter out noisy formation. Meanwhile, GRU cells are computed faster than LSTM cells. After obtaining the hidden vector h_t, h_c, we use them to calculate the topic attention weight h_a as follows:

$$
\begin{aligned}
h_{a_i} &= (\mathbf{W}_{a_i} h_t + b_{a_i}) \odot h_c \\
h_a &= (h_{a_1}, h_{a_2}, h_{a_3}, ..., h_{a_m})
\end{aligned}
$$

In the combination-level layer, the h_o is concatenated by the hidden vectors of h_c, h_z (topic-similarity vector), h_t (topic-id vector) and h_a (utterance-topic attention), then h_o is forwarded to the softmax as follows:

$$
I(\mathbf{c}^{-r}, r) = softmax(\mathbf{W}_o h_o + b_o)
$$

where $\mathbf{W}_o$ and b_o are the parameters. To learn all these parameters, we minimize the cross entropy as follows:

$$
-\sum (y_i \log I(c_i^{-r}, r_i) + (1 - y_i) \log(1 - I(c_i^{-r}, r_i)))
$$

The MHCN architecture well integrates the word-level and topic-level information of utterances. Moreover, the MHCN model regards the topical similarity between sentences and the similarity between sentence embeddings as well, which makes the model robust to capture the topical coherence between utterances.

4 EXPERIMENTS

In this section, we present the experimental results[1]. In Section 4.1, we evaluate the proposed framework in terms of Precision, Recall and F1. In Section 4.2, we demonstrate the impact of model parameters, followed by a case study in Section 4.3.

4.1 Data set and Baseline

We first briefly describe the data sets[2] used in our study. Then several the state-of-the-art baselines are introduced. Finally, we report the performance of cpLDA and MHCN for the task of discourse coherence.

Ubuntu Dialogue Corpus (UDC)[3] is derived from the Ubuntu Relay Chat Channel [17]. This dataset contains about 1.85 million conversations with an average of 5 utterances per conversation.

Weibo. To learn the cpLDA model, we introduce a Weibo data set, which consists of 1.1 million status-comment pairs. We treat each status-comment pair as a single-turn conversation and each status or comment as one sentence. We filter out the single term and stopwords in the data set. After preprocessing, the vocabulary size of the Weibo data set is 33,067.

Baidu Tieba. We construct our dialogue data set from Baidu Tieba, which serves as the largest Chinese forum and allows users to post status or comment on any topic. Our Tieba data consists of 1 million dialogues with multiple turns. We treat the current utterance as the response one and all the previous utterances in the same dialogue as the context. We regard the response in the dialogue as the positive one, and randomly sample the negative response from the Tieba data.

Baselines. We present two types of baselines: discriminative models and generative models.

- **Discriminative models** include *LSTM*, *CNN* [12], *Hierarchical RNN (HRNN)* [15], *MLP* and *HGRU* [15].

[1] In order to enhance reproducibility, the source code of the proposed framework is provided in the supplementary material.
[2] The data sets will be made publicly accessible upon publication.
[3] http://www.iulianserban.com/Files/UbuntuDialogueCorpus.zip

Table 3: Results of Discourse Coherence Evaluation

Methods	UDC			Tieba		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
LSTM	0.7912	0.7961	0.7905	0.6230	0.6245	0.6229
CNN	0.8275	0.8263	0.8278	0.6253	0.6268	0.6250
HRNN	0.8317	0.8401	0.8309	0.5830	0.5841	0.5817
MLP	0.8304	0.8302	0.8273	0.6212	0.6222	0.6205
HGRU	0.8392	0.8415	0.8394	0.6421	0.6441	0.6404
SEQ2SEQ	0.7528	0.7591	0.7513	0.5530	0.5589	0.5517
SEQ2SEQ-MMI	0.7532	0.7410	0.7408	0.6023	0.6014	0.5996
VLV-GM	0.7986	0.7973	0.7869	0.6506	0.6503	0.6498
MHCN	**0.8969**	**0.8930**	**0.8938**	**0.7341**	**0.7319**	**0.7284**
MHCN(SLDA)	0.8691	0.8656	0.8634	0.6644	0.6652	0.6635
w/o H-cpLDA	0.8355	0.8368	0.8376	0.6533	0.6541	0.6516
w/o H-topic	0.8241	0.8244	0.8225	0.6482	0.6478	0.6459
w/o HGRU	0.7953	0.7917	0.7933	0.6129	0.6105	0.6146
w/o Attention	0.8740	0.8719	0.8682	0.7103	0.7114	0.7062

- **Generative models** employed in our study are *SEQ2SEQ* [1], *SEQ2SEQ-MMI* [14] and *VLV-GM* [14]. *SEQ2SEQ* is an encoder-decoder model with the attention mechanism to predict the current utterance given its context. *SEQ2SEQ-MMI* can be regarded as the advanced *SEQ2SEQ* model with the objective function Maximum Mutual Information (MMI). As reported in [14], *SEQ2SEQ-MMI* yields more diverse responses compared with *SEQ2SEQ*. *VLV-GM* is a model that combines encoder-decoder and a Markov chain using global information z_n like LDA. *VLV-GM* is closely related to many attempts in training variational autoencoders(VAE) [13], variational or latent-variable recurrent nets [5, 6], hierarchical latent variable encoder-decoder models [17].

Experimental Setup. For all models, we set the dimensionality of word vectors to 200. We adopt the stochastic gradient descent (SGD) and the learning rate is set to 0.0005. The batch size is 1024 and the epoch times is 200. All the models are implemented using tensorflow and trained on a Tesla P40 GPU. For the cpLDA and sentenceLDA, we experimentally set the topic number to 200 and the epoch times to 400. The cpLDA and sentenceLDA are learned from the Weibo data set and the UDC data, and then applied to infer the topic of utterance for the Tieba data set and the UDC data.

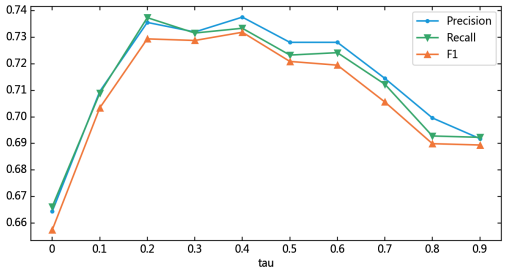
4.2 Results Analysis and Parameter Sensitivity

Results Analysis. Table 3 demonstrates the experimental results. We could observe that our MHCN model outperforms all state-of-the-art models measured in terms of all metrics. In order to investigate the influence of different components of our model, we also replace cpLDA with SLDA which is short for sentenceLDA. In specific, we discard hierarchical cpLDA (noted as w/o H-cpLDA), hierarchical topic embedding (noted as w/o H-topic), hierarchical GRU (noted as w/o HGRU) and attention mechanism in combination-layer (noted as w/o Attention), respectively.

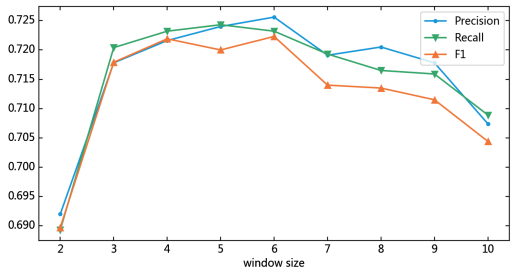
We summarize the observation as follows: (1) *SEQ2SEQ* yields the worst results measured by precision, indicating that the generative models could not well distinguish positive responses from negative ones. One possible reason is that the generative models are more likely to yield safe or simple response, which can not match the real response well. (2) *MHCN* achieves better performance than both *HGRU* and *MHCN(SLDA)*. It proves that the mined topic embeddings facilitate the performance of the *MHCN* model. Moreover, Compared with *MHCN(SLDA)*, *MHCN* achieves significant improvements in terms of all metrics, which justifies the capability of cpLDA on generating fine-grained topics. (3) In *MHCN* achieves, each parts make contributions to the final performance. In specific, the attention mechanism find out the common information of topic representations and utterance representations. Therefore it makes topic and utterance representations combine more closely.

Parameter Sensitivity. We proceed to investigate the influence of model parameters. First, we tune the trade-off parameter τ for global coherence in cpLDA (see Eq. 1). Recall that τ is introduced to balance the local and global coherences of topics, and a larger τ indicates that the cpLDA is more influenced by the global coherence. We vary τ from 0.0 to 0.9 with the step 0.1 and train the cpLDA models, then the resultant topics inferred from the corresponding cpLDA are fed to the MHCN model to measure the performance of discourse coherence. The experimental results are reported in Figure 3(a). We observe that τ slightly influences the merits (i.e., precision, recall and F1) when τ varies from 0.1 to 0.9. In addition, the MHCN model with $\tau \in [0.2, 0.4]$ achieves the best performance. However, when τ is set to 0.0, meaning that only local coherence is considered (same to the sentenceLDA model), the performance measured by any merit drops sharply. This observation shows the importance of the global coherence in cpLDA.

Next, we investigate the parameter window size W on the model performance. W represents the number of utterances in each discourse. Recall that in Section 4.1 we set the window size to the fixed number 6. Here we slide the window on the discourse and select number of utterances ranging from 2 to 10. In another word, we evaluate the coherence between $W-1$ utterances and the current utterance. The experimental results are shown in Figure 3(b). We discover that the MHCN model with $W \in (3, 4, 5, 6)$ achieves the best performance, and then slightly declines.



(a) Varying Trade-off Parameter τ



(b) Varying Window Size W

Figure 3: Parameter Sensitivity

	Chinese	English	Score		
			HGRU	VLV-GM	MHCN
Coherent Cases	A: … B: … A: 别睡了,快点起来开工了 B: 俺好累好困啊 A: 每天睡十一小时	A: … B: … A: Don't sleep, get up to work! B: I am so tired and sleepy. A: You almost sleep for 11 hours everyday!	0.0210	0.2989	0.9943
	A: … B: … A: 偶是射手座射手挺多的样子 B: 射手较花心 A: 可是专出美女帅哥喔	A: … B: … A: I' m a Sagittarius, it seems that so many Sagittarius here! B: Sagittarius are always faithless! A: But so many handsome boys and pretty girls are Sagittarius.	0.1075	0.3661	0.8570
Incoherent Cases	A: … B: … A: 甜点好吃,使我无法自拔 B: 这个看着就好吃啊 A: 超级惊人的饭量	A: … B: … A: The dessert tastes good, and I like it very much! B: It looks very delicious! A: Gigantic appetite.	0.6452	0.8760	0.3405
	A: … B: … A: 她年龄太小 B: 年龄不是问题 A: 这是一个数学问题	A: … B: … A: She is too young. B: Age is not a problem. A: This is a mathematical problem.	0.9888	0.5921	0.0184

Figure 4: Case Study

Finally, the phenomenon indicates that τ and W as the value increases, the performance first increases and then decreases. One possible reason is that the more information is provided, the more noise is introduced.

4.3 Case Study

We show several cases in Figure 4. Each case contains 3 utterances (one current utterance and two contextual utterances) and coherence scores generated by different methods. The coherence score is calculated as the coherence degree between the current utterance and its context. We quantitatively compare our *MHCN* with the best discriminative baseline *HGRU* and the best generative baseline *VLV-GM*. For the first case on topic "sleep", the current utterance "You almost sleep for 11 hours everyday!" is topically coherent with its two previous utterances (i.e., "Don't sleep, get up to work!" and "I am so tired and sleepy."). While *VLV-GM* and *HGRU* yield lower scores 0.2989 and 0.0210. In contrast, *MHCN* obtains much larger coherence score 0.9943. Similar observations can been seen in the second case on topic "constellation". For the third case, the current utterance "gigantic appetite." is obviously irrelevant with its two previous utterances on topic "dessert". Only *MHCN* detects the incoherence between utterances and assigns the smallest score. Similar result is observed for the last case. Due to the page limination, we only show these four cases. In practice, we analyze a massive amount of cases, and find that, compared with the state-of-the-art discriminative and generative baselines, our neural structure with the fed topic embedding is able to capture the semantic relation between utterances more precisely. That is, the topic information of utterances is well encoded and decoded in *MHCN*, facilitating the computation of the coherence degree between utterances. Our *MHCN* model is superior to other state-of-the-art models in the discourse coherence task for the dialogue system.

5 CONCLUSION

In this paper, we tackle with the problem of discourse coherence. The main objective of this research is to distinguish the coherent utterances from incoherent ones given the context. To address the problem, we propose a novel framework that seamlessly integrates Bayesian and neural components. These two components are Coherence-Pivoted Latent Dirichlet Allocation (cpLDA) and

Multi-Hierarchical Coherence Network (MHCN). The former aims to infer fine-grained topics by considering both local and global semantics coherence. The latter employs multiple GRU cells and takes full advantage of the topics generated by cpLDA. Experimental results show that the proposed framework achieves superior performance comparing to several state-of-the-art methods.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).

[2] David M. Blei and John D. Lafferty. 2005. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18*. 147–154.

[3] David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Machine Learning, Proceedings of the Twenty-Third International Conference*. 113–120.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[5] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating Sentences from a Continuous Space. *Computer Science* (2015).

[6] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. 2015. A Recurrent Latent Variable Model for Sequential Data. *Computer Science* 35, 8 (2015), 1340–1353.

[7] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. 601–610.

[8] Miroslav Dudík, David M. Blei, and Robert E. Schapire. 2007. Hierarchical maximum entropy density estimation. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference*. 249–256.

[9] Micha Elsner, Joseph L. Austerweil, and Eugene Charniak. 2007. A Unified Local and Global Model for Discourse Coherence. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*. 436–443.

[10] Camille Guinaudeau and Michael Strube. 2013. Graph-based Local Coherence Modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. 93–103.

[11] Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining*. 815–824.

[12] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. *Eprint Arxiv* 1 (2014).

[13] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. (2013).

[14] Jiwei Li and Dan Jurafsky. 2017. Neural Net Models of Open-domain Discourse Coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 198–209.

[15] Zhao Meng, Lili Mou, and Zhi Jin. 2017. Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection. (2017).

[16] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[17] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. (2016).

[18] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. LightLDA: Big Topic Models on Modest Computer Clusters. In *Proceedings of the 24th International Conference on World Wide Web*. 1351–1361.

Learning to Represent the Evolution of Dynamic Graphs with Recurrent Models

Aynaz Taheri
University of Illinois at Chicago
Chicago, Illinois
ataher2@uic.edu

Kevin Gimpel
Toyota Technological Institute at Chicago
Chicago, Illinois
kgimpel@ttic.edu

Tanya Berger-Wolf
University of Illinois at Chicago
Chicago, Illinois
tanyabw@uic.edu

ABSTRACT

Graph representation learning for static graphs is a well studied topic. Recently, a few studies have focused on learning temporal information in addition to the topology of a graph. Most of these studies have relied on learning to represent nodes and substructures in dynamic graphs. However, the representation learning problem for entire graphs in a dynamic context is yet to be addressed. In this paper, we propose an unsupervised representation learning architecture for dynamic graphs, designed to learn both the topological and temporal features of the graphs that evolve over time. The approach consists of a sequence-to-sequence encoder-decoder model embedded with gated graph neural networks (GGNNs) and long short-term memory networks (LSTMs). The GGNN is able to learn the topology of the graph at each time step, while LSTMs are leveraged to propagate the temporal information among the time steps. Moreover, an encoder learns the temporal dynamics of an evolving graph and a decoder reconstructs the dynamics over the same period of time using the encoded representation provided by the encoder. We demonstrate that our approach is capable of learning the representation of a dynamic graph through time by applying the embeddings to dynamic graph classification using a real world dataset of animal behaviour.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Learning latent representations**; *Unsupervised learning*.

KEYWORDS

dynamic graph; representation learning; recurrent models

ACM Reference Format:
Aynaz Taheri, Kevin Gimpel, and Tanya Berger-Wolf. 2019. Learning to Represent the Evolution of Dynamic Graphs with Recurrent Models. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316581

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316581

1 INTRODUCTION

Dynamic graphs are a popular model for high level representation, characterization, and analysis of real world dynamic systems. Interactions among the unit entities of a dynamic system are typically modeled and represented by a dynamic graph or a dynamic network, also referred to as a temporal graph [21, 22]. A dynamic graph is a graph or a network in which the edges between the nodes change over time, and nodes can appear and disappear from the network during the network evolution. Dynamic graphs are often defined as time-ordered sequences of network snapshots, and each network snapshot models the interactions between the nodes over a unit interval of time.

Graph representation learning approaches gained significant attention in downstream graph analysis applications in recent years. Several studies focused on learning to represent nodes and edges in dynamic graphs [16, 29, 39]. However, the ultimate goal of these approaches is still far from capturing both the topological and temporal features of entire time-varying graphs. Recently, remarkable achievements have been obtained for representation learning of whole static graphs [1, 27, 35, 40], but a comprehensive fundamental framework for processing dynamic graphs is still lacking. There are numerous differences between static and dynamic graphs that cause methods for embedding static graphs to not be applicable to dynamic graphs. The possibility of information diffusion through a series of time-varying graphs is a complex process and depends on the temporal dynamics of the graphs. Unlike static graphs, temporal dynamics, not just the topology of the connections, is the major determinant of information flow in dynamic graphs. Therefore, an ideal approach should be able to learn about the temporal dynamics in order to represent a dynamic graph.

Our goal is to demonstrate the effectiveness of our approach for dynamic graph representation learning that is able to learn the topology of the graph as well as the temporal dynamics of the graph. To that end, we propose an unsupervised approach for combining the power of static graph representation methods and the success of recurrent neural networks in learning sequences of events. The proposed model empowers the representation learning approaches to formulate the dynamic graph representation learning through gated graph neural networks (GGNNs) [25], and recurrent neural networks. Our unsupervised dynamic graph representation approach can be used not only in processing labeled data, such as in dynamic graph classification, but can be also applied in many practical applications, such as anomaly detection in temporal social networks or streaming data, as well as in exploratory analysis and scientific hypothesis generation. Our approach is constructed based on a sequence-to-sequence architecture that learns the dynamic

graph evolution. GGNNs are embedded in a recurrent encoder to preserve the topology of a dynamic graph at each time step. In addition, long short-term memory networks (LSTMs) are incorporated to capture the dynamics by propagating temporal information between consecutive time steps. An autoregressive decoder is leveraged to reconstruct the history of graph evolution from the encoder hidden representation. To the best of our knowledge, this paper presents the first approach in learning to represent the entire graph in a dynamic setting. We demonstrate the efficacy of our approach in a dynamic classification task using a real dataset of baboon behaviour. We provide several baselines to demonstrate the efficacy of our approach in the classification task and present quantitative analysis to indicate the effects of considering dynamics in our representation learning method.

2 RELATED WORK

We briefly discussed some of the prior related work in section 1. In this section, we discuss the static and dynamic graph representation learning approaches in more detail.

2.1 Static Graphs

Recently, there has been a burst of methods for embedding nodes [17, 31] and subgraphs [1, 27, 40] of static graphs into low-dimensional spaces. In this paper, we only focus on the related work for representation learning of whole graphs. Graph2vec [28] is an unsupervised method inspired by document embedding models. This approach finds a representation for a graph by maximizing the likelihood of graph subtrees given the graph embedding. However, Graph2vec still does not quite capture the global information in the graph structure since it is limited to only considering subtrees as graph representatives. Taheri *et al.* [35] proposed an unsupervised architecture using sequence-to-sequence encoder-decoder models to capture the entire graph structure.

Several approaches [11, 12, 23, 25, 30, 41] have been proposed to learn graph representations for a specific supervised task, such as graph classification or regression. The representations obtained by these approaches are tailored for a supervised task and are not determined solely based on the graph structure. Niepert *et al.* [30] developed a framework (PSCN) to learn graph representations using convolutional neural networks (CNNs). Deep Graph Convolutional Neural Network (DGCNN) [41] is another model that applies CNNs for graph classification. The main difference between DGCNN and PSCN is the way they deal with the vertex-ordering problem. Lee *et al.* [23] leveraged recurrent neural networks and an attention mechanism to find informative parts of the graph for the purpose of graph classification. Message passing neural networks [12] are another group of supervised approaches that have been recently used for graph structured data, such as molecular property prediction in chemistry [11, 12, 25]. Duvenaud *et al.* [11] introduced a CNN to create "fingerprints" (vectors that encode molecule structure) for graphs derived by molecules. The information about each atom and its neighbors are fed to the neural network, and neural fingerprints are used to predict new features for the graphs. Bruna *et al.* [3] proposed spectral networks, generalizations of CNNs on low-dimensional graphs via graph Laplacians. Henaff *et al.* [18] and Defferrard *et al.* [8] extended spectral networks to high-dimensional graphs. Scarselli *et al.* [33] proposed graph neural networks (GNN) and find node representations using random walks. Li *et al.* [25] extended GNNs with gating recurrent neural networks, GGNNs, to predict sequences from graphs. Generally, GGNNs have been used in the supervised tasks in the literature. However, our architecture is an unsupervised framework in order to capture the temporal and topological aspects of a dynamic graph.

2.2 Dynamic Graphs

Lately, there have been a few studies on representation learning for dynamic graphs. The majority of these studies focused on representation learning for individual nodes within the dynamic graphs. Nguyen *et al.* [29] suggested using temporal random walks and the skip-gram model for learning node embeddings. Also, Du *et al.* [9] proposed an extension of the skip-gram model in a dynamic setting to learn the representation of new nodes and adjust the old ones. Trivedi *et al.* [36] presented an architecture, Know-Evolve, for node representation learning in a temporal knowledge graph using recurrent neural networks and temporal point processes. DyRep [37] extended Know-Evolve with a two-time scale process that captures temporal node interactions in addition to the topological evolution. Sankar *et al.* [32] proposed a framework for node representation learning including two consecutive self-attention modules for aggregation of structural and temporal information respectively. DyGEM [16] incrementally computes node representations by initializing an autoencoder from the previous step. Goyal *et al.* [15] extended DyGEM by adding recurrent neural autoencoders in order to capture more accurate temporal information of node embeddings. DynamicTriad [42] imposed triad to learn the node embeddings while preserving the temporal information. Chen *et al.* [5] modified LSTMs in order to process dynamic networks and predict the upcoming links in the future. Wang *et al.* [38] introduced a small-scale method for representation learning of a series of transition graphs in the area of driving analysis. The approach is tuned for processing small-size graphs with a fixed number of nine nodes, and the entire adjacency matrix of the graph is flattened to a vector and used as the initial representation of the graph and input of the system. However, this approach does not scale to processing large-size, sparse, and variable number of nodes through time. In contrast, our approach focuses on the entire graph representation learning and preserves both topological and temporal properties of a graph.

3 PROBLEM DEFINITION

Dynamic graph: a dynamic graph is an ordered sequence of T graph snapshots: $G = \langle G_1, G_2, \ldots, G_T \rangle$. Graph $G_t = (V_t, E_t)$ models the state of a dynamic system at the interval $[t, t + \Delta t]$, for some fixed Δt. The dynamic graph G may have a subset of nodes from the set V at each time step, $V_t \subset V$. Each node $v \in V_t$ takes a unique identification value from $1, \ldots, |V|$, and edge $e^t_{kj} \in E_t$ is a pair of nodes $(k, j) \in \{E_t : V_t \times V_t\}$ and represents an edge at time step t. The initialization of the embedding (node representation) for node v is denoted by $x_v \in \mathbb{R}^d$. The initialization of the node embedding is explained in Section 7. G_t is represented by an adjacency matrix $A_t \in \mathbb{R}^{n \times n}$, where n is the number of nodes and where $a^t_{kj} = 1$ if

there is an edge between nodes k and j at time step t, and $a^t_{kj} = 0$ otherwise.

Timestep history: a sequence of graphs seen before time step t is called the history of G_t and represented by: $H_{G_t} = \langle G_{t-w}, G_{t-w+1}, G_{t-w+2}, \ldots, G_{t-1} \rangle$. H_{G_t} indicates a window that we have observed in the past w time steps in order to learn the graph evolution.

Graph embedding: we seek to learn a mapping function $\Phi(G_t|H_{G_t})$ that embeds a graph G_t into $\mathbb{R}^k$. The goal is to learn graph embeddings such that graphs having similar topological structure and temporal dynamics lie close to one another in the embedding space. We focus on processing undirected graphs in this work, but it is trivial to extend our models to directed graphs.

4 PRELIMINARIES AND NOTATION

Graph Neural Networks (GNNs) [14, 33] are a well-known neural network architecture for processing graph structures, and the main idea is to learn the topology of a graph via an iterative propagation procedure. These methods are sometimes referred to as Message Passing Neural Networks (MPNNs) [13] in the literature. This architecture learns node representations using a message propagation method and then aggregates the representation to find an embedding for the entire graph. The message propagation procedure runs for several steps, and updates the hidden states of the nodes at each iteration. Finally, a representation is computed for the whole graph at the last iteration based on those hidden states.

Gated Graph Neural Networks (GGNNs). Li et al. [25] proposed GGNNs as an extension of GNNs in order to learn the reachability across the nodes in a graph. They used Gated Recurrent Units (GRU) [6] to propagate a message from a node to all its reachable nodes. This allows backpropagation to unroll the recurrence in several steps and causes the hidden states of all reachable nodes to be updated. In the following we formally define the propagation model in GGNNs. Let A denote the input adjacency matrix of a graph and we will use A_v to denote row v of that matrix. Let x_v denote the initial embedding of a node v, which can be obtained in many ways and we will discuss the initialization used in this work in Section 7. We use n^v_i to indicate the hidden state of a node v at iteration i. With this notation, we have the following propagation model.

We initialize the hidden states of each node n^v_0 at the beginning as follows:

$$n^v_0 = x_v \tag{1}$$

Each propagation step passes information from the neighbors of a node v to learn its embedding a^v_i at propagation step i:

$$a^v_i = A_{v:}[n^1_{i-1} \ldots n^{|V|}_{i-1}] + b \tag{2}$$

where b is a bias. Note that after several iterations of this step the information is passed among all the reachable nodes to learn the embedding of each node.

The remaining equations are the update (Equation 3) and reset (Equation 4) gates of the GRU, that update the nodes' hidden states by incorporating information from the previous iterations of the propagation step. The matrices W, U are the parameter matrices for the GRU, σ is a logistic sigmoid activation function, and tanh is a hyperbolic function to add non-linearity to produce the first estimate $\widetilde{n}^v_i$ of an embedding of a node v at step i. (The symbol $\odot$ denotes elementwise matrix multiplication.) The final embedding n^v_i of a node v at step i is given by Equation 6:

$$\begin{aligned}
z^v_i &= \sigma(W^z a^v_i + U^z n^v_{i-1}) && (3)\\
r^v_i &= \sigma(W^r a^v_i + U^r n^v_{i-1}) && (4)\\
\widetilde{n}^v_i &= \tanh(W a^v_i + U(r^v_i \odot n^v_{i-1})) && (5)\\
n^v_i &= (1 - z^v_i) \odot n^v_{i-1} + z^v_i \odot \widetilde{n}^v_i && (6)
\end{aligned}$$

Long short-term memory (LSTM). We provide a brief overview of LSTMs [20]. An LSTM is a recurrent neural network (RNN) designed to model long-distance dependencies in sequential data. We denote the input vector at time t by g_t and we denote the hidden vector computed at time t by h_t. At each time step, an LSTM computes a memory cell vector c_t, an input gate vector i_t, a forget gate vector f_t, and an output gate vector o_t:

$$\begin{aligned}
i_t &= \sigma(W_i g_t + U_i h_{t-1} + K_i c_{t-1} + b_i)\\
f_t &= \sigma(W_f g_t + U_f h_{t-1} + K_f c_{t-1} + b_f)\\
o_t &= \sigma(W_o g_t + U_o h_{t-1} + K_o c_{t-1} + b_o) && (7)\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c g_t + U_c h_{t-1} + b_c)\\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

Here, again, $\odot$ denotes elementwise multiplication, σ is the logistic sigmoid function, each W is a weight matrix connecting inputs to particular gates (denoted by subscripts), each U is an analogous matrix connecting hidden vectors to gates, each K is a diagonal matrix connecting cell vectors to gates, and each b is a bias.

Encoder-Decoder: In this paper, we use a sequence-to-sequence framework to train our encoder-decoder model. We refer to Equation 7 as an "LSTM encoder" because it is the main step that converts an input sequence into a sequence of hidden vectors h_t. We will also use a type of LSTM that predicts the next item $\overline{x}$ in the sequence from h_t. This architecture, which we refer to as an "LSTM decoder," adds the following:

$$\overline{g} = f(h_t) \tag{8}$$

where f is a function that takes a hidden vector and outputs a predicted observation $\overline{g}$. With symbolic data, f typically computes a distribution over symbols and returns the most probable symbols. When using continuous inputs, f could be an affine transform of h_t followed by a nonlinearity.

5 RECURRENT MODEL FOR DYNAMIC GRAPHS

Given the standard notation and existing methodology described in section 4 and the problem definition in section 3, we describe our architecture to learn the embedding of a dynamic graph $\Phi(G_t|H_{G_t})$ in this section. The proposed architecture is composed of a recurrent encoder that projects the timestep history H_{G_t} into a k dimensional space. GGNNs are applied to model the graph topology at each time step, and more specifically provide the input for the encoder. The recurrent structure of the encoder allows us to embed the G_t using the hidden representation of the timestep history H_{G_t} and the graph topology at time step t.

DyGGNN: The model includes three main components: 1) a *GGNN* to capture the topology of the graph at time step t, 2) an *encoder*

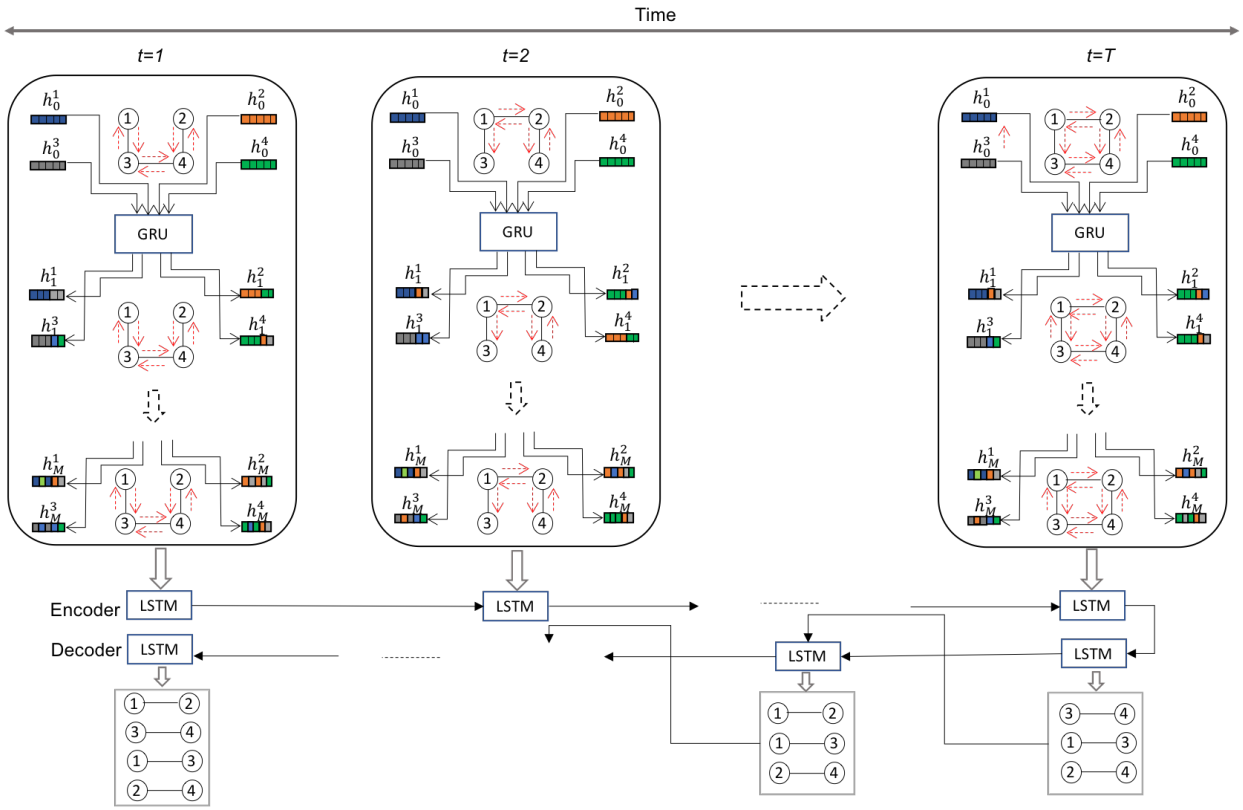


Figure 1: DyGGNN model schematic. A GGNN models the graph topology at each time step and provides the input to the LSTM encoder. The hidden representation of a sequence of graphs is used by the decoder to reconstruct the dynamic graph evolution.

that projects the graph evolution over a window of T time steps into a k dimensional space, and 3) a *decoder* which reconstructs the structure of the dynamic graph at each time step using its hidden representation. Figure 1 shows the architecture of our approach for representation learning of a dynamic graph over T time steps. In the following, we explain the components of this architecture in detail.

The GGNN builds a graph representation for G_t by considering its topological structure at time step t. After M steps of message propagation in G_t, we use an average pooling of the nodes' hidden states (Equation 6) as the representation of the entire G_t, and we refer to this as a static embedding of the graph:

$$Emb_{st}(G_t) = Avg(h_M^v)$$

Given the static embedding of the graph $Emb_{st}(G_t)$ and its history H_{G_t}, we use an LSTM encoder to project G_t into a hidden representation that takes the graph dynamics into account. The LSTM encoder, $LSTM_{enc}$, passes the information of the dynamic graph G over an observation window of size w and computes the dynamic embedding G_t using knowledge about the graph topology from the past w time steps.

$$h_t^{enc} = LSTM_{enc}(Emb_{st}(G_t), h_{t-1}^{enc}) \tag{9}$$

We refer to h_t^{enc} as the dynamic embedding of the graph, $Emb_{dy}(G_t)$. The main goal of the decoder, $LSTM_{dec}$, is to reconstruct the history of the graph H_{G_t} in the observed window with size w. The decoder is as an autoregressive model and predicts the topology of the graph at time step t given the recently predicted graphs at previous time steps. The decoder uses h_T^{enc} to initialize its first hidden state. The set of edges of the graph is the prediction of the decoder at each time step. The f function in Equation 8 is instantiated as follows:

$$h_t^{dec} = LSTM_{dec}(\overline{Edges}(G_{t-1}), h_{t-1}^{dec}) \tag{10}$$

We use a sigmoid transformation σ to predict the existence/nonexistence of an edge at time step $t-1$ (here W are the learned parameters and b is a bias):

$$\overline{Edges}(G_{t-1}) = \sigma(h_{t-1}^{dec} * W + b) \tag{11}$$

6 TRAINING

Let O be a set of temporal windows where each window $o \in O$ has a length of w. The parameter space of the models includes parameters related to $GGNN$, $LSTM_{enc}$, $LSTM_{dec}$, $W \in \mathbb{R}^{|h^{dec}| \times |V|^2}$ and $b \in \mathbb{R}^{|V|^2}$. We use a cross-entropy loss function to train our model:

$$Loss_{CE} = \sum_{o \in O} \sum_{t=1}^{|o|} \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} A_{ij}^{t} \log(\overline{Edges}(G_t)_{ij})) \quad (12)$$

In order to train an efficient model, we need to pay attention to the choice of the temporal window size, w, since we define the connectedness of the temporal graph over the length of the window. There may be some very long temporal paths that cannot be captured with a window size of w. On the other hand, we cannot consider very long windows in practice, because we face the gradient vanishing problem during backpropagation. Some approaches [10, 19] truncate very long sequences of individual events to a batch of shorter independent sequences and use backpropagation though time over the mini-batches of those sequences. However, in our task we have a dynamic graph G which is a sequence of graph snapshots, and we need to take care of the sequence of nodes' hidden states. We cannot consider the sequences of nodes independently because we need to consider both temporal and topological structure of the graph. Trivedi et al. [36] suggested using a different type of mini-batches for event detection on edges of a temporal graph. They select k sequential events through time and made a batch with size k from the events. Therefore, at each mini-batch, only parameters related to nodes and edges constructing those k events are updated. Their approach cannot fully utilize the history of nodes, topology, and dynamics of a temporal graph. Using a sliding observation window, we decompose the dynamic graph into several dynamic graphs. The sequential windows are grouped together to make a mini-batch (Algorithm 1). In this way, we preserve not only the temporal dependency at each window, but also we capture the implicit relations between sequential windows, since the model parameters are updated considering the whole mini-batch input data.

Algorithm 1 Mini-batch Backpropagation on Temporal Graphs

Input: Dynamic graph $G : (G_1, G_2, \ldots, G_T)$, Window size w, Batch size b, Sliding gap s
$cur_{win} = 0, b_{ind} = 0, minbatch = \{\}$
while $cur_{win} < T$ **do**
  while $b_{ind} < b$ **do**
    $sample = G[cur_{win} : cur_{win} + w]$
    $minibatch.Add(sample)$
    $b_{ind} = b_{ind} + 1$
    $cur_{win} = cur_{win} + s$
  end while
  Use $minibatch$ to find the $\overline{Edges}$ (Eq. 11)
  Minimize loss function (Eq. 12) through w time steps
  $minibatch = \{\}, b_{ind} = 0$
end while

7 EXPERIMENTS

In this section, we evaluate our dynamic representation learning approach using our proposed model. We use our learned representation for a dynamic graph classification task. At each time step t, the graph G_t has a label l from the set L. Our unsupervised learned representations for G_t are used in the classification task.

7.1 Dataset

We use dynamic graphs of a troop of GPS-tracked baboons living in the wild in Mpala Research Centre, Kenya [7, 34]. The adult and sub-adult members of the troop were fitted with GPS collars, collecting data points at 1Hz for 12 hours (6am to 6pm) for 28 days. The data are a 2-day subset of the dataset from 16 adult and sub-adult members of the troop. The subset was chosen due to the availability of the activity labels, provided by biologists and extended by sequence classification approaches [24]. Amornbunchornvej *et al.* [2] constructed dynamic graphs of individuals following the behavior patterns of each other and performed the activity classification task. The dynamic graph of the first day has 23259 time steps and the second day has 19098 time steps. There are four group level activities in the baboon dataset: sleeping, hanging-out, coordinated non-progression, and coordinated progression. One of the four labels is assigned to the dynamic graph at each time step by the domain experts. We report two results, following the configuration baselines in [2, 24]: using the labeled first day of data to classify the second day's activities (Baboon (day 1)); and using the second day's labeled data to classify the first day's activities (Baboon (day 2)).

7.2 Baselines

We compare our results with several baselines:

- Li *et al.* [24] proposed an Adversarial Sequence Tagging (AST) to perform activity labeling classification for the baboon dataset.
- Amornbunchornvej *et al.* [2] used dynamic graph features at each time step to predict the activity label using a linear discriminant analysis (LDA).
- We also compare with other baselines that learn static graph representations. Taheri *et al.* [35] proposed an unsupervised sequence-to-sequence (S2S) approach for static graph representation learning. We use this approach to find the representation of G_t without considering its history.
- Moreover, we use a supervised GGNN classifier as a baseline to classify G_t based on the graph topology at the same time step.

7.3 Hyperparameters

The GGNN uses four iterations of message passing at each time step. The dimensionality of the GRU and LSTMs is fixed to 100. The Adam optimizer is used for minibatch training. We set the learning rate to be 0.001 and dropout rate to 0.5. A unique identifier is assigned to each baboon in the troop. We represent identifiers via one-hot vectors to initialize the x_v in GGNNs. Figure 2 indicates the effect of window size and batch size on the classification performance. It shows that the best accuracy is obtained by larger window sizes and batch sizes, which points to the fact that longer-term dynamics matter for this task. Having larger window sizes causes an explicit larger context for representation learning during backpropagation, and having larger batch sizes causes an implicit larger context through time. Smaller window sizes are more sensitive to the batch

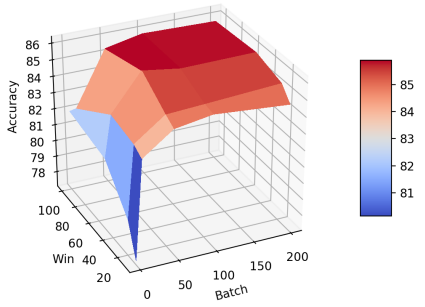


Figure 2: The effect of window size and batch size on accuracy.

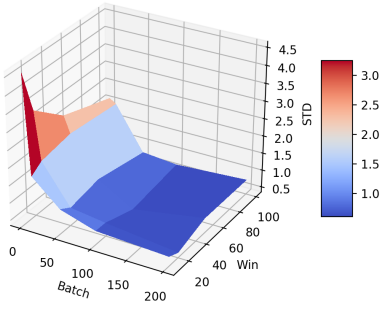


Figure 3: The effect of window size and batch size on standard deviation.

size, and small batch sizes do not perform well especially with small window sizes. Figure 2 shows that we get the best results with window sizes larger than 50 and batch size larger than 50. We fixed the batch size and window size to 50 in the rest of our experiments. The standard deviations of different configuration of window sizes and batch sizes are shown in Figure 3. Each point in the plot shows the standard deviation of 10 experiments. Larger batch sizes have an important role in the robustness of our experiments. Increasing the batch size provides smaller standard deviations while smaller window sizes are prone to greater standard deviations when the batch size is small.

7.4 Results

We use a C-SVM classifier from LIBSVM [4] with a radial basis kernel for multi-class classification. The data of one day is used for training and tuning the regularization and kernel hyperparameters of the SVM via cross-validation and the other day is used for the test. Each experiment is repeated 10 times and we report average accuracies and standard deviations. We compare our proposed models with other baselines in Table 1. we exceed all other baselines and the performance shows that dynamics matter and incorporating temporal information in embedding methods improves graph representations.

Table 1: Activities classification prediction accuracy in first two days of baboon data

Method	Baboon Day 1	Baboon Day 2
AST	77.30	69.22
LDA	87.20	70.82
GGNN	84.58±1.2	81.38±1.1
S2S	84.66±1.3	83.87±1.5
DyGGNN	88.02±1.8	86.21±1.7

7.5 Model analysis

Our model computes representations for G_t from two points of view: A static representation of the topology of the graph at time step t: $Emb_{st}(G_t)$ and a dynamic representation of the graph: $Emb_{dy}(G_t)$. Figure 4 shows the difference between the representations found by the $Emb_{st}(G_t)$ and $Emb_{dy}(G_t)$ functions. The accuracy at time step t in the plot shows the accuracy of the model in the interval $[0 \dots t]$. The generated representations from $Emb_{dy}(G_t)$, which include the dynamic notion of graph evolution, outperform the embeddings obtained by the $Emb_{st}(G_t)$, which consider only the topology of the graph at time step t. Furthermore, Figure 4 highlights that training the model through time improves the performance of the dynamic embedding significantly in comparison with a static representation.

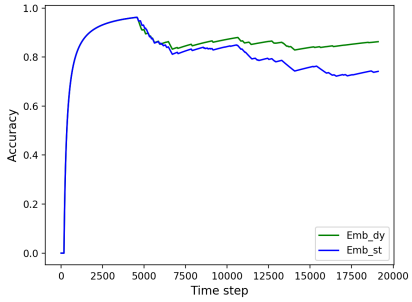


Figure 4: The classification accuracy of the model in the interval of $[0 \dots t]$ for the two embeddings of $Emb_{st}(G_t)$ and $Emb_{dy}(G_t)$.

Figure 5 shows the representation of the baboon dataset using t-SNE [26]. Each point shows the graph assigned to one of the time steps. The four class labels were not used when learning the graph representation, but we still find the graphs are well clustered. Since these are human observed activities and there is a natural transition in behavior (these are not discrete behaviors), there is a continuity of labels in the representations. We actually do not expect *a priori* cluster separation. Note, that most of the activity is "hanging out", which is biologically consistent. The sleeping activity is seen at the bottom center of the right figure in Figure 5 and can be thought of as a special kind of "hanging out". The "coordinated

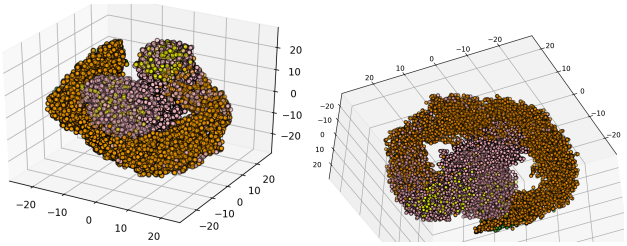


Figure 5: Visualization of the baboon dataset in a 3-dimensional space (Orange: Hanging-out, Pink: Coordinated progression, Yellow: Coordinated non-progression, Green: Sleeping)

non-progression" is a transitional activity from “hanging-out" to “coordinated progression”.

8 CONCLUSIONS

We propose an efficient approach for representation learning of entire graphs in a dynamic setting. We evaluate our approach in a real-world dataset of animal behaviour and show that our model achieves significantly better performance in comparison with state-of-the-art models that learn the graph representation in a static setting. The promising results confirm our approach as a solid foundation for dynamic graph analysis.

REFERENCES

[1] B. Adhikari, Y. Zhang, N. Ramakrishnan, and B. A. Prakash. 2017. Distributed Representations of Subgraphs. In *DaMNet*.

[2] Chainarong Amornbunchornvej, Ivan Brugere, Ariana Strandburg-Peshkin, Damien R. Farine, Margaret C. Crofoot, and Tanya Y. Berger-Wolf. 2018. Coordination Event Detection and Initiator Identification in Time Series Data. *ACM Trans. Knowl. Discov. Data* 12, 5 (2018), 53:1–53:33.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral Networks and Locally Connected Networks on Graphs. *CoRR* (2013).

[4] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM TIST* 2 (2011).

[5] Jinyin Chen, Xuanheng Xu, Yangyang Wu, and Haibin Zheng. 2018. GC-LSTM: Graph Convolution Embedded LSTM for Dynamic Link Prediction. *arXiv preprint arXiv:1812.04206* (2018).

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] MC Crofoot, RW Kays, and Wikelski M. 2015. Data from: Shared decision-making drives collective movement in wild baboons. *Movebank Data Repository* (2015).

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv* (2016).

[9] Lun Du, Yun Wang, Guojie Song, Zhicong Lu, and Junshan Wang. 2018. Dynamic Network Embedding: An Extended Approach for Skip-gram based Network Embedding.. In *IJCAI*.

[10] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1555–1564.

[11] David Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*.

[12] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. *CoRR* (2017).

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212* (2017).

[14] Marco Gori, Gabriele Monfardini, and Franco Scarselli. [n. d.]. A new model for learning in graph domains. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks (IJCNN'05)*, Vol. 2. IEEE, 729–734.

[15] Palash Goyal, Sujit Rokka Chhetri, and Arquimedes Canedo. 2018. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. *arXiv preprint arXiv:1809.02657* (2018).

[16] Palash Goyal, Nitin Kamra, Xinran He, and Yan Liu. 2017. DynGEM: Deep Embedding Method for Dynamic Graphs. *CoRR* abs/1805.11273 (2017).

[17] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *KDD*.

[18] M. Henaff, J. Bruna, and Y. LeCun. 2015. Deep convolutional networks on graph-structured data. *arXiv* (2015).

[19] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *ICLR*.

[20] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[21] Petter Holme and Jari Saramäki. 2012. Temporal networks. *Physics reports* 519, 3 (2012), 97–125.

[22] David Kempe, Jon Kleinberg, and Amit Kumar. 2002. Connectivity and inference problems for temporal networks. *J. Comput. System Sci.* 64, 4 (2002), 820–842.

[23] John Boaz Lee, Ryan Rossi, and Xiangnan Kong. 2018. Graph classification using structural attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1666–1674.

[24] Jia Li, Kaiser Asif, Hong Wang, Brian D Ziebart, and Tanya Y Berger-Wolf. 2016. Adversarial Sequence Tagging.. In *IJCAI*.

[25] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2016. Gated graph sequence neural networks. In *ICLR*.

[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* (2008).

[27] Annamalai Narayanan, Mahinthan Chandramohan, Lihui Chen, Yang Liu, and Santhoshkumar Saminathan. 2016. subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs. *MLG* (2016).

[28] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning Distributed Representations of Graphs. In *MLG*.

[29] Giang Hoang Nguyen, John Boaz Lee, Ryan A Rossi, Nesreen K Ahmed, Eunyee Koh, and Sungchul Kim. 2018. Continuous-time dynamic network embeddings. In *International Workshop on Learning Representations for Big Networks at The Web Conference*.

[30] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning Convolutional Neural Networks for Graphs. In *ICML*.

[31] B. Perozzi, R. Al-Rfou, and S Skiena. 2014. DeepWalk: Online learning of social representations. In *KDD*.

[32] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. 2018. Dynamic Graph Representation Learning via Self-Attention Networks. *arXiv preprint arXiv:1812.09430* (2018).

[33] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2009), 61–80.

[34] Ariana Strandburg-Peshkin, Damien R Farine, Iain D Couzin, and Margaret C Crofoot. 2015. Shared decision-making drives collective movement in wild baboons. *Science* 348, 6241 (2015), 1358–1361.

[35] Aynaz Taheri, Kevin Gimpel, and Tanya Berger-Wolf. 2018. Learning Graph Representations with Recurrent Neural Network Autoencoders. In *KDD Deep Learning Day*.

[36] Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. 2017. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *ICML*.

[37] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. 2019. DyRep: Learning Representations over Dynamic Graphs. In *International Conference on Learning Representations*.

[38] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Pengfei Wang, Yu Zheng, and Charu Aggarwal. 2018. You are how you drive: Peer and temporal-aware representation learning for driving behavior analysis. In *KDD*. ACM.

[39] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.

[40] Pinar Yanardag and SVN Vishwanathan. 2015. Deep graph kernels. In *KDD*.

[41] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. 2018. An End-to-End Deep Learning Architecture for Graph Classification. In *AAAI*.

[42] Le-kui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. 2018. Dynamic Network Embedding by Modeling Triadic Closure Process.. In *AAAI*.

Spectral Graph Complexity

Anton Tsitsulin
University of Bonn

Davide Mottin
Aarhus University

Panagiotis Karras
Aarhus University

Alex Bronstein
Technion

Emmanuel Müller
University of Bonn

ACM Reference Format:
Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alex Bronstein, and Emmanuel Müller. 2019. Spectral Graph Complexity. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316589

1 INTRODUCTION

Graphs conveniently model relationships, interactions, physical and semantic structures. On such graphs, data mining tasks, such as community detection and classification, are performed either directly or on low-dimensional embeddings [12, 18] derived from the graph structures. By its nature, a graph represents complex interactions among objects in an abstract manner, while the underlying data intrinsically possess a notion of complexity.

Unfortunately, such complexity is hard to quantify and assess, since the inherent shape of the data is unknown. As such, there current way to have an implicit indication of the data complexity emerges after an analyst has engaged a graph in an involved data mining task. Yet, the result of an analysis can be inexplicably bad or even meaningless, leaving no other choice than repeating the experiment without a precise indication whether said experiment will fail or succeed with different settings. To avoid pointless analyses, we necessitate a characterization of the intrinsic complexity of the graph structure upfront.

Several works have tried to characterize some aspects of a graph's structure [4, 9] such as degree distribution or community structure. Yet, none of them provides a characterization of the intrinsic complexity of general graphs. Besides, recent efforts from the fields of chemistry and biology [13, 15] have been constrained to a domain-specific context, hindering the applicability to few specific cases [7, 16].

We take a different approach and study the graph from the lens of its spectrum. Spectral graph theory [5] studies the properties of a graph based on the eigenvalues of the graph's Laplacian. The full spectrum provides insights that allow for defining similarities between graphs [17], and performing various different tasks, such as spectral clustering [19] or community detection [10]. While the properties of the first non-zero eigenvalue are well understood especially in the context of algebraic connectivity [1, 6], the full characterization of the spectrum is still under investigation. The Spectral properties of networks have been extensively used in quantum gravity studies, mainly in connection to synchronization properties [2]. According to Weyl's Law [21], the growth rate of the Laplacian eigenvalues is inversely proportional to the dimensionality of an underlying manifold. The recently introduced complex network manifold model [3] uses d-dimensional simplicial complexes as the building block for constructing more complex graphs with tunable spectral dimension. Millan et al. [11] study synchronization properties of the complex network manifold model with relation to its spectral dimension. However, a characterization of graphs in terms of spectral dimensionality has not been proposed before.

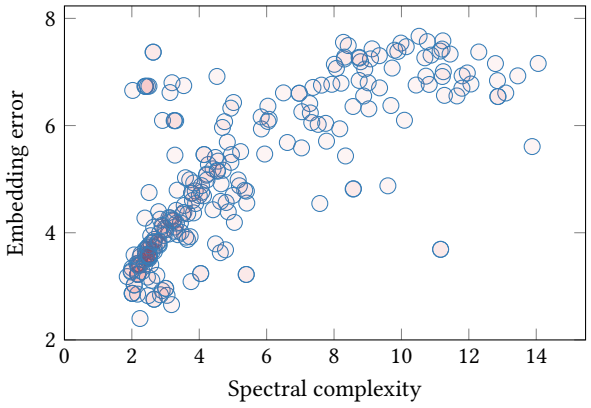


Figure 1: Quality of VERSE embedding [18] correlates to our spectral complexity measure. Spearman correlation is 0.76, while mutual information equals 0.82.

In this paper, we introduce the *first*, to our knowledge, universal complexity measure for general graphs which is directly deduced from the spectrum of graphs. We show a preliminary result that demonstrates that the spectral dimension reflects the underlying complexity of real-world graphs by correlating spectral complexity measure to embedding quality on a large collection of real-world graphs.

2 SPECTRAL COMPLEXITY ESTIMATION

We study the complexity of undirected graphs. A graph is a pair $G = (V, E)$, where $V = (v_1, \ldots, v_n), n = |V|$ is the set of vertices and $E \subseteq (V \times V)$ the set of edges. We assume the graph has no weights on edges even though our approach readily applies to the weighted case.

The *adjacency matrix* of a graph G is a $n \times n$ matrix A having $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise. A graph's *normalized Laplacian* is the matrix $\mathcal{L} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is the diagonal matrix with the degree of node i in the entry D_{ii}, i.e,

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316589

$D_{ii} = \sum_{j=1}^{n} A_{ij}$. Since the Laplacian is a symmetric matrix, its eigenvectors $\phi_1, \ldots, \phi_n$, are real and orthogonal to each other. Thus, it allows eigendecomposition as $\mathcal{L} = \Phi\Lambda\Phi^\top$, where Λ is a diagonal matrix on the sorted eigenvalues $\lambda_1 \leq \ldots \leq \lambda_n$ of which $\phi_1, \ldots, \phi_n$ are the corresponding eigenvectors, and Φ is an orthogonal matrix obtained by stacking the eigenvectors in columns $\Phi = [\phi_1 \phi_2 \ldots \phi_n]$.

The set of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ is called the *spectrum* of a graph. The normalized Laplacian, as opposed to the unnormalized version $L = D - A$, has a bounded spectrum, $0 \leq \lambda_i \leq 2$. In general, the normalized Laplacian has more attractive theoretical properties than its unnormalized counterpart [20].

The normalized Laplacian matrix has the same spectrum as the random walk Laplacian $L^{\text{rw}} = I - D^{-1}A$, which is closely related to the propagation of random walks in the graph. A random walk process starts from any node in the graph and at each steps randomly select one of the neighbor nodes. The random walk process can be studied in terms of its probability of occurrence of a random walker at node i at time t by the following equation:

$$\pi_i(t) = -\sum_j L^{\text{rw}}_{ij} \pi_j(t-1).$$

The *spectral dimension* is commonly defined [8] as the asymptotic behavior (at large times) of the probability of a random walk to return to the starting point $\pi_G(t) \sim t^{-d_S/2}, t \to \infty$. More formally,

$$d_S = -2 \lim_{t \to \infty} \frac{\log \pi_G(t)}{\log t}.$$

However, this definition is useless for characterizing the complexity of finite graphs, as for every finite graph $d_S = 0$ under this formulation, since the return probability is always 1. As real graphs are finite, we can leverage the fact that the spectral dimension can be estimated from the eigenvalue growth rate $\rho(\lambda)$ [21]:

$$\rho(\lambda) \simeq \lambda^{d_S/2} \tag{1}$$

for $\lambda \ll 1$. In d-dimensional Euclidean lattices $d_S = d$. More generally, the spectral dimension relates to the Hausdorff (fractal) dimension which intuitively accounts for the local distances between points at multiple scales:

$$d_H \geq d_S \geq 2\frac{d_H}{d_H + 1}$$

However, as far as we know, no result similar to the Weyl's law for manifolds has been obtained for finite graphs. We take the analogy to the discrete case and build our spectral dimensionality estimator based on Weyl's law.

Given a graph G with n vertices, we first compute its full spectrum or an approximation $\{\lambda_k\}$. We then linearly interpolate the spectrum producing a line $\lambda(x)$ within the interval $[0, 1]$ such that $\lambda(k/n) = \lambda_k$. The interpolated spectrum $\lambda(x)$ is then sampled on a fixed grid $(x_1, \ldots, x_M)$ with M points, producing an M-dimensional vector $\tilde{\lambda}$ with entries $\tilde{\lambda}_k = \lambda(x_k)$. The vector $\tilde{\lambda}$, having fixed size, is insensitive to the graph size and also invariant to the ordering of its vertices. Finally, we select a point s (in our experiments $s = 1/100$) and estimate the slope of the initial part of the spectrum $\tilde{\lambda} \leq \lambda(s)$. This spectrum slope approximates the asymptotic growth of the graph's eigenvalues.

3 PRELIMINARY EXPERIMENT

We downloaded 259 networks with size 10^3 to 10^5 from the Network Repository [14], computed the spectrum of their normalized Laplacians, and estimated the spectral dimension as outlined in the Equation 1. Next, we computed VERSE embeddings[18], using the default personalized PageRank (PPR) similarity, 128 dimensions, and $c = 0.85$. We evaluate the capacity of VERSE of predicting the PPR similarity performance in terms of KL divergence with the real PPR vectors.

Figure 1 shows strong relationship between our estimation of the spectral complexity and the objective performance on a graph task. While there is some indication that not all networks are scale-free [4], we observe this phenomenon in terms of the spectral dimension, as only 3 out of 259 graphs have spectral dimension less than 2 which is related to scale-free networks.

As spectral dimensionality indicates the "hardness" of the graph for the embedding algorithms and embeddings are closely related to a variety of downstream tasks [18], we envision important results in the understanding of data mining tasks on graphs. Our preliminary result attests the importance of the study of complexity measures for graph and defends the choice of spectral analysis as a privileged tool for the understanding of the performance of graph tasks. This study of infinite graphs provides rich insights for the finite case.

REFERENCES

[1] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. 2016. Exact Recovery in the Stochastic Block Model. *IEEE Transactions on Information Theory* (2016).
[2] Alex Arenas, Albert Díaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou. 2008. Synchronization in complex networks. *Physics reports* (2008).
[3] Ginestra Bianconi and Christoph Rahmede. 2015. Complex quantum network manifolds in dimension d> 2 are scale-free. *Scientific reports* 5 (2015), 13979.
[4] Anna D. Broido and Aaron Clauset. 2018. Scale-free networks are rare. (2018).
[5] Fan RK Chung. 1997. *Spectral graph theory*. Number 92. American Math. Soc.
[6] Nair Maria Maia de Abreu. 2007. Old and new results on algebraic connectivity of graphs.
[7] Matthias Dehmer, Nicola Barbarini, Kurt Varmuza, and Armin Graber. 2009. A large scale analysis of information-theoretic network complexity measures using chemical structures. *PLoS One* 4, 12 (2009), e8057.
[8] Bergfinnur Durhuus. 2009. Hausdorff and spectral dimension of infinite random graphs. *Acta Phys. Polon.* 40 (2009).
[9] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the internet topology. In *ACM SIGCOMM review*.
[10] Kyle Kloster and David F Gleich. 2014. Heat kernel based community detection. In *KDD*. 1386–1395.
[11] Ana P Millán, Joaquín J Torres, and Ginestra Bianconi. 2018. Synchronization in Network Geometries with Finite Spectral Dimension. *arXiv:1811.03069* (2018).
[12] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *kDD*. 701–710.
[13] Douglas Rigterink and David J Singer. 2014. A network complexity metric based on planarity and community structure. *Journal of Complex Networks* (2014).
[14] Ryan A. Rossi and Nesreen K. Ahmed. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. http://networkrepository.com
[15] Kaushik Sinha and Olivier L de Weck. 2016. Matrix Energy as a Measure of Topological Complexity of a Graph. *arXiv preprint arXiv:1608.08456* (2016).
[16] Sudarshan Sridhar, Mohammad Fazelpour, Amaninder S Gill, and Joshua D Summers. 2016. Accuracy and precision analysis of the graph complexity connectivity method. *CIRP* 44 (2016), 163–168.
[17] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander M. Bronstein, and Emmanuel Müller. 2018. NetLSD: Hearing the Shape of a Graph. In *KDD*.
[18] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. 2018. VERSE: Versatile Graph Embeddings from Similarity Measures. In *WWW*.
[19] Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17 (2007), 395–416.
[20] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. 2008. Consistency of spectral clustering. *The Annals of Statistics* (2008).
[21] Hermann Weyl. 1911. Über die asymptotische Verteilung der Eigenwerte. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen* (1911).

FdGars: Fraudster Detection via Graph Convolutional Networks in Online App Review System

Jianyu Wang*
xiaoyu_paopao@zju.edu.cn
College of Computer Science,
Zhejiang University
Hangzhou, China

Rui Wen*
rachelwen@tencent.com
Platform and Content Group, Tencent
Shenzhen, China

Chunming Wu
wuchunming@zju.edu.cn
College of Computer Science,
Zhejiang University
Hangzhou, China

Yu Huang
neilyuhuang@tencent.com
Platform and Content Group, Tencent
Shenzhen, China

Jian Xiong
janexiong@tencent.com
Platform and Content Group, Tencent
Shenzhen, China

ABSTRACT

Online review system enables users to submit reviews about the products. However, the openness of Internet and monetary rewards for crowdsourcing tasks stimulates a large number of fraudulent users to write fake reviews and post advertisements to interfere the rank of apps. Existing methods for detecting spam reviews have been successful but they usually aims at e-commerce (e.g. Amazon, eBay) and recommendation (e.g. Yelp, Dianping) systems. Since the behaviors of fraudulent users are complex and varying across different review platforms, existing methods are not suitable for fraudster detection in the online app review system.

To shed light on this question, we are among the first to analyze the intentions of fraudulent users from different review platforms and categorize them by utilizing characteristics of content (similarity, special symbols) and behaviors (timestamps, device, login status). With a comprehensive analysis of spamming activities and relationships between normal and malicious users, we design and present FdGars, the first graph convolutional network approach for fraudster detection in the online app review system. Then we evaluate FdGars on a real-world large-scale dataset (with 82,542 nodes and 42,433,134 edges) from Tencent App Store. The result demonstrates that F_1-score of FdGars can achieve 0.938+, which outperforms several baselines and state-of-the-art fraudsters detecting methods. Moreover, we deploy FdGars on Tencent Beacon Anti-Fraud Platform to show its effectiveness and scalability. To the best of our knowledge, this is the first work to use graph convolutional networks for fraudster detection in the large-scale online app review system. It is worth to mention that FdGars can uncover malicious accounts even the data are lack of labels in anti-spam tasks.

*Both authors contributed equally to this research. The student author finished this work during intern.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316586

CCS CONCEPTS

• **Information systems** → **Web indexing**; **Spam detection**; • **Hardware** → Dynamic memory.

KEYWORDS

Fraud Detection, Graph Convolutional Networks, Online App Review System

ACM Reference Format:
Jianyu Wang, Rui Wen, Chunming Wu, Yu Huang, and Jian Xiong. 2019. FdGars: Fraudster Detection via Graph Convolutional Networks in Online App Review System. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316586

1 INTRODUCTION

Online review systems provide services for people to share their opinions and make decisions, such as which clothes to buy and where to eat. It is usually regarded as a type of explicit feedback signal for products. Since users can generate content containing rich information, online review system has become an attractive target for fraudsters.

Detecting fake reviews and reviewers is a non-trivial problem. Prior works [11, 13, 14, 17, 19, 20] attempt to solve this problem by using language models, analyzing abnormal behaviors and building graphs to discover suspicious patterns. However, they mostly focus on detecting spammers in e-commerce platform. Few research has been done to detect multi-class fraudsters in different review systems. Figure 1 shows the different review patterns between Yelp and Myapp Store. Fraudsters in Yelp aim to influence consumers' decisions by writing masses of reviews with high quality, always camouflage like normal reviewers. In contrast, fraudsters in application stores, also called spammers, aim to post advertisements (e.g. telephone number, Wechat, URL) to promote the rank of apps (e.g. flushing, crowdturfing) and so on.

In this paper, we investigate the characteristics of spam reviews from different platforms, including online marketplace (eBay, Amazon), app store (Myapp, Google Play) and content service (Yelp,

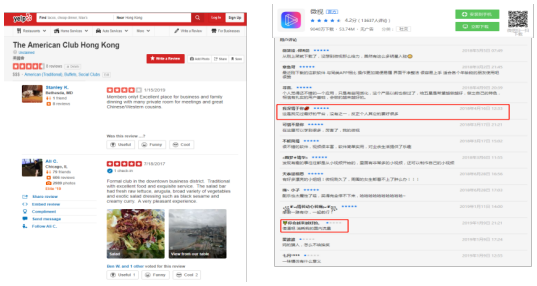


Figure 1: Comparison Between Different Review Platform.

Dianping). In addition, we also categorize fraudsters by their motivations such as spreading fake information, promoting product ranking and advertising into 3 types:

- Camouflage [6]: fraudsters pretend themselves as normal reviewers by adding links to popular items or famous people;
- Crowdturfing [8]: fraudsters can easily hire web workers to take part in particular spam activities with monetary award from crowdsourcing platforms (e.g. Rapidworkers, Microworkers).
- Spammer: fraudsters publish irrelevant reviews in open review systems to increase popularity of their products (e.g. advertisements, product information) or do some illegal activities (e.g. selling drugs, sensitive words).

As fraudsters in review platforms are adversarial, irregular and distributed, anti-spam tasks face a huge challenge. In the field of fraud detection, many works focus on the characteristics of content. They develop text or Nature Language Processing (NLP) models to distinguish fake reviews from legitimate reviews [13]. Since users' behavioral attributes (e.g. timestamps, footprints) can yield clues as to which are fraudulent, behavior-based approaches also attract many researchers and are widely applied in many industrial works [11]. Recently, graph-based detection methods [1, 2, 5, 6, 19, 20, 20] have been studied with inspiration from network embedding approaches, which leverage the relationship between users, reviews and items, and made considerable progress in spotting malicious accounts.

In order to detect fraudsters in large-scale app stores, we propose a method named FdGars by combining reviewers' features and Graph Convolutional Network (GCN). Firstly, we analyze review logs and extract content and behavior features for each users. Secondly, a graph structure is constructed to express the characteristics of reviewers and relationships between reviewers. Then, reviewers are classified into fraudsters and normal users through a predefined labeling method. Based on the limited labeled reviewers, a two-layer GCN is developed to detect more fraudsters from unlabeled reviewers.

Through experiments, we evaluate FdGars over the sample of 82,542 reviewers and 302,097 reviews. As shown in Figure 2, the constructed graph includes 82,542 nodes and 42,433,134 edges. We label the reviewers on 1 Aug, 2018 as training set and detect more fraudsters from 2 Aug, 2018 and 3 Aug, 2018. Finally, FdGars's recall reaches to 0.958+. To make the result more convincing, LR, RF and DeepWalk are used and their recalls are 0.516+, 0.828+ and 0.911+ respectively, demonstrating the effectiveness of the presented method.

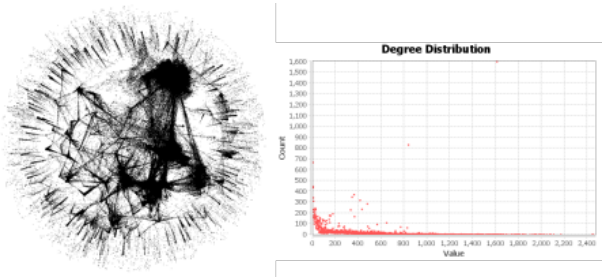


Figure 2: User graph in online App review platform.

We summarize the contributions of this work as follows:

1. *Fraudsters Analysis*. We shed light on detecting different types of fraudster in online review systems and conclude 3 popular patterns of generating fake/spam reviews by fraudsters.
2. *FdGars Implementation*. We develop an efficient and scalable methods system by using graph convolutional networks for anti-spam tasks in a large-scale online app review system. FdGars performs better than other state-of-the-art methods for detecting malicious accounts.
3. *Deployment and Evaluation*. We validate the performance of FdGars in a real word dataset by deploying it on the Tencent's Beacon Anti-Fraud Platform. Our method achieves both high precision and high recall.

2 METHODOLOGY

In this section, we first give an overview of FdGars, followed by the fraudster measurement and GCN.

2.1 FdGars Framework

The procedure of FdGars is illustrated in Figure 3.

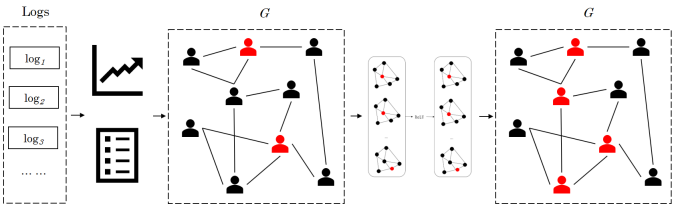


Figure 3: Diagram of FdGars Framework.

First of all, we extract content features and behavior features for each reviewer based on their review logs. Content feature concentrates on a reviewer's text content, such as a review's symbol ratio and the quantity of similar reviews. Behavior feature focuses on a reviewer's behavior in a designated period, including the review quantity, the 24-hours distribution of review quantity and the score distribution.

Second, we construct a graph structure G according to the following rules:

1. represent reviewers as nodes in graph G;
2. construct an edge to connect two nodes if their corresponding reviewers have reviewed the same app.

Consequently, the review logs are transformed into a graph structure. The graph G can be expressed as

$$G = \{\mathbf{N}, \mathbf{E}, \mathbf{A}\} \tag{1}$$

Where, $\mathbf{N} = \{n_1, n_2, n_3, ...\}$ is the reviewer set, $\mathbf{E} = \{e_1, e_2, e_3, ...\}$ is the edge set and $\mathbf{A} = \{a_1, a_2, a_3, ...\}$ is the reviewer attribute set. An edge between nodes n_i and n_j indicates that reviewers n_i and n_j have reviewed the same app, denoted as

$$e_{ij} = Edge(n_i, n_j) = (n_i, a_i, n_j, a_j) \tag{2}$$

After the construction, graph G clearly expresses the characteristics of reviewers and relationships between reviewers.

Third, we utilize some rules to label high suspicious fraudsters. As described in Section 2.2, reviewers are divided into fraudsters and normal users. We design a labeling method to label high suspicious fraudsters and normal users. According to our data analysis, we believe the labeling method can recognize high suspicious fraudsters to some extent.

Then, we train a GCN model to learn node feature and graph structure according to graph G and a small account of labeled reviewers. The GCN's detail is introduced in Section 2.3. When the training is finished, we utilize the learned model to find more fraudsters from unlabeled reviewers in app store. Experiments demonstrate the performance of the proposed FdGars method.

2.2 Fraudsters Measurement

In reality, accessing fraudsters is a hard work since malicious accounts always change their strategies to avoid detection systems. Traditional methods for labeling fraudsters usually relay on researcher opinion or a team of human labors, but they have obvious limitations by just utilizing single characteristics such as language feature, posting times and the frequency of posting reviews to make evaluations. In contrast to previous efforts, we aims to collect fraudsters as seeds who have strong evidence to anti-spam tasks at first by leveraging reviews' content features and behavior distributions. Then we propagate these seeds to find more malicious accounts, who have strong connections with seed users.

Review Classification. Firstly, we use content features to classify the reviews. Table 1 shows the features we extract from each review.

Table 1: Content Features and Description

Feature	Description
SRN	Similar Review Number
RSN	Review Symbol Number
RL	Review Length
PRR	Proportion of RSN to RL
REB	Regular Expression Blacklist

- **SRN** is the number of similar reviews in a designate period, which calculated by *SimHash* algorithm ;
- **RSN** is the number of special character in each review, such as emoji expression and Mars symbols. In general, spam review contains more than 50% symbols;
- **RL** is the length of each review;
- **PRR** is the ratio of symbol number and review length;
- **REB** is the collection of regular expression for reviews posting by spammers with obvious intentions including telephone number ($\hat{}1(3|4|5|7|8)\backslash d\{9\}\$$), ambiguous word ($[WwVv][Xx]$), URLs and so on.

Reviewer Analysis. Secondly, we classify reviewers by utilizing two behavior attributes: the continuous days and the number of login device.

- **CD** is the continuous days that reviewers have posted reviews in a designated period;
- **DN** records the device quantity that reviewers have used in the same period.

We design a labeling function expressing as

$$Label(n) = \begin{cases} 1 & if\ CD > \theta_{CD}\ \&\ DN > \theta_{DN} \\ 0 & otherwise \end{cases} \tag{3}$$

where, θ_{CD} and θ_{DN} are thresholds. If an reviewer's CD > θ_{CD} or DN > θ_{DN}, we consider the reviewer as a fraudster; otherwise, the reviewer is a normal user.

Through the multiple rounds of data analysis, we believe $\theta_{CD} = 7$ and $\theta_{DN} = 20$ can recognize high suspicious fraudsters in the scenario of online app store. Table 2 lists the features we extract for each reviewer:

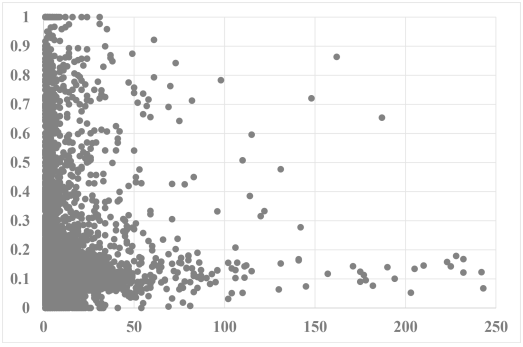
Table 2: Behavior Features and Descriptions

Feature	Description	Latitude
RQ	Review Quantity	1
TQD	Time-based Quantity Distribution	24
SQD	Score-based Quantity Distribution	5

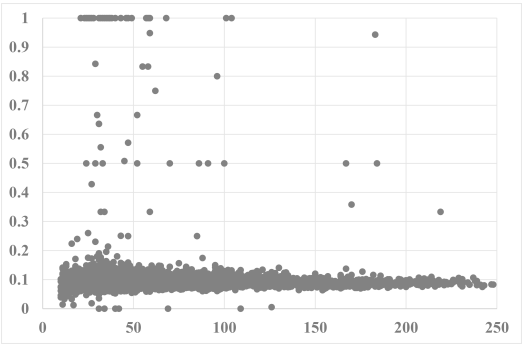
- **RQ** records the quantity of an reviewer's reviews in a designated period.
- **TQD** is the 24-hours quantity distribution of an reviewer's reviews.
- **SQD** is an reviewer's rating distribution.

By labeling reviewers based on Equation 3, Figure 4 shows the fraudster and normal user distributions with RQ and PRR. Obviously, normal users gather at the bottom-left of the coordinate. Based on our analysis, The RQ of most normal users are less than 30. Comparatively, fraudsters gather at the bottom of the coordinate. Most fraudsters' PRRs are less than 0.2. Consequently, $\theta_{CD} = 7$ and $\theta_{DN} = 20$ can be used to recognize high suspicious fraudsters and label reviewers in the following experiment.

Fraudsters Identification. Since our data set is not labeled by human labors, we use review classification results to make sure that our ground truth is necessarily an approximation. We evaluate each reviewers by considering the number of spam reviews he/she has posted during a month. In this way, we can identify fraudsters with high confidence even their reviews or behaviors look normal.



(a) Normal Users



(b) Fraudsters

Figure 4: Distribution comparison between Fraudsters and Normal Users with RQ and PRR features. (Note: X-axis represents RQ and Y-axis represents PRR.)

2.3 Graph Convolutional Networks

We use a two-layer GCN for semi-supervised fraudster detection in this paper. Take a graph G in Figure 5 as an example, nodes in G represent 6 reviewers and edges represent their behavioral relationships.

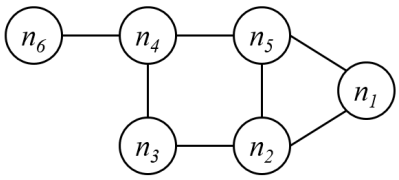


Figure 5: A Simple Example of Graph G.

First, we calculate a symmetric adjacency matrix A on graph G. According to Figure 5, the result of A is

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Second, we get a matrix X to express reviewers' feature vectors which are described in Section 2.2. Then, we define a two-layer GCN as the following layer-wise propagation rule:

$$H^{(0)} = X$$

$$H^{(1)} = \sigma[\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(0)}W^{(0)}]$$

$$H^{(2)} = \sigma[\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(1)}W^{(1)}]$$

Where, $H^{(1)}$ is the matrix of activations in the $1st$ layer and $H^{(2)}$ is the matrix of activations in the $2nd$ layer. $\widetilde{A} = A + I_N$ is the adjacency matrix of the undirected graph with added self-connections. I_N is the identity matrix. D is defined as

$$D = \begin{bmatrix} \sum_{j=0}^{N-1}\widetilde{A}_{0,j} & 0 & \dots & 0 \\ 1 & \sum_{j=0}^{N-1}\widetilde{A}_{1,j} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & \sum_{j=0}^{N-1}\widetilde{A}_{N-1,j} \end{bmatrix}$$

$W^{(0)}$ is an input to hidden weight matrix for a hidden layer with H feature maps. $W^{(1)}$ is a hidden to output weight matrix. Both of them are layer-specific trainable weight matrices. $\sigma(\cdot)$ denotes an activation function, such as $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$.

Finally, we train the model $f(X, A)$ according to the steps described above and predicted results will be the outputs of $f(X, A)$, expressing as

$$f(X, A) = \mathrm{softmax}\Big(\hat{A}\,\mathrm{ReLU}\Big(\hat{A}XW^{(0)}\Big)W^{(1)}\Big)$$

Here, the softmax activation function is defined as

$$\mathrm{softmax} = \frac{1}{\mathcal{Z}}exp(x_i)$$

where $\mathcal{Z} = \sum_i \exp(x_i)$.

In conclusion, the two-layer GCN is schematically depicted in Figure 6. It is constructed based on a small amount of labeled reviewers as Section 2.2 described. After training, the learned model can detect more fraudsters who have similar behaviors with the labeled fraudsters from unlabeled reviewers.

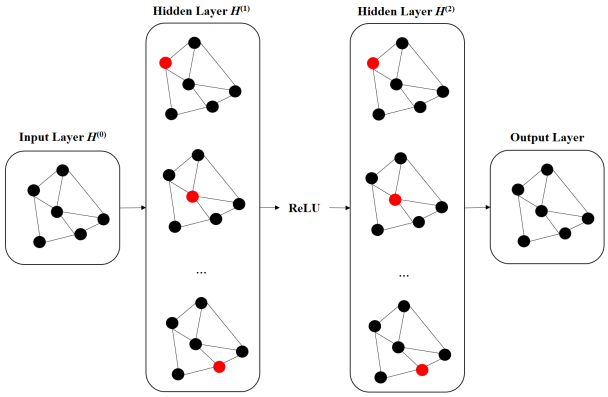


Figure 6: Graph Convolutional Network.

3 EXPERIMENTS

In this section, the presented detection method is evaluated on real world big data and deployed on Tencent's Venus Computation Platform.

3.1 Data Collection and Annotation

To validate the performance of the proposed method, large-scale review datasets are utilized from Tencent Inc. and its detail is listed in Table3. It can be seen that 82,542 users have posted 302,097 reviews from 1 Aug, 2018 to 3 Aug, 2018.

Table 3: Review Data Statistics

Date	Users	Apps	Reviews
1 Aug, 2018	31,450	4,374	94,378
2 Aug, 2018	32,540	4,328	105,958
3 Aug, 2018	35,104	4,559	101,761
All Three Days	82,542	7,584	302,097

We take the reviewers on Aug 1, 2018 as training set and the reviewers on 2 Aug and 3 Aug as testing set. To verify the diffusion performance of the presented fraudster detection, one in particular is that our testing set has excluded the reviewers who have posted reviews in training set. Table 4 shows the graph information based on our dataset. It is worth mentioning that the graph construction process costs nearly 2 minutes and the ratio of positive reviews is 38.1%.

Table 4: Graph Statistics

Nodes	Edges	Label	Quantity
82,542	42,433,134	1	5,926
		0	25,524

3.2 Baseline Methods

We compare the performance of FdGars with two widely-used methods for classification in many fields. We also add two famous graph structure based methods as our baseline methods.

- **Logistic Regression** According to the features listed in Table 2 and the labeling dataset mentioned in Table 3, we train a Logistic Regression (LR) classifier. Then, we use the classifier to divide the reviewers into fraudsters and normal users from the unlabeled reviewers.
- **Random Forest** Similarly, we train a Random Forest (RF) classification model. The learned RF model is also used to find fraudsters from the unlabeled reviewers.
- **DeepWalk** We train a DeepWalk [16] model to learn latent representations of nodes in our graph. In our experiment, the number of random walks is set to 10; the length of random walk is 80; the embedding length is 128. At last, we train a RF model to test DeepWalk's performance.
- **LINE** We train a LINE [18] model to learn latent representations of nodes in our graph. The negative ratio is set to 5 and the embedding length is 128. At last, we train a RF model to test LINE's performance.

3.3 Performance Evaluation

Table 5 lists the training and predicting time costs. It is worth mentioning that the training and predicting procedures are executed on CPU. According to our statistics, the predicting time lasts nearly 14 minutes.

Table 5: Time Costs of Training and Predicting Procedure

Epochs	Training Time	Predicting Time
100	39.37 minutes	13.65 minutes
500	150.58 minutes	12.83 minutes

Table 6 shows the predicted results with different epochs, as are True Positive (TP), False Positive (FP), True Negative (TN) and False Positive (FP). To measure the performance of the proposed method, Table 7 shows the details of metrics, as are $Accuracy$, $Recall$, $Precision$ and F_1-score.

Table 6: Evaluation Results with Different Epochs

Epochs	Date	TP	FP	TN	FN
100	2 Aug, 2018	1,335	205	23,520	572
	3 Aug, 2018	2,444	347	24,849	601
500	2 Aug, 2018	1,776	327	23,398	131
	3 Aug, 2018	2,913	587	24,609	132

Table 7: Performance Statistics with Different Epochs

Epochs	Date	$Recall$	$Precision$	F_1
100	2 Aug, 2018	0.7001	0.8669	0.7746
	3 Aug, 2018	0.8026	0.8757	0.8376
500	2 Aug, 2018	0.9588	0.9195	0.938
	3 Aug, 2018	0.9667	0.9135	0.9379

Table 8 presents the performances of fraudster detection with different methods. Obviously, LR and RF have lower recalls comparing with other methods. It is noteworthy that DeepWalk takes 170.17 minutes to calculate each node's embedding and LINE spends almost 116 hours. Comparatively, FdGars only takes 150.58 minutes as expressed in Table 5. This indicates that FdGars has a better computational efficiency in real applications.

Figure 7 shows the ROC curves with different datasets. We also test FdGars on real world review logs from 1 Sep, 2018 to 3, Sep 2018. By labeling reviewers in 1 Sep, 2018 as training set, we calculate the detection results from new reviewers in 2 Sep, 2018 and 3 Sep, 2018. The four curves indicate that FdGars can maintain a stable performance for fraudster detection.

Table 8: Performance Comparison with Baseline Methods

Method	Date	*Recall*	*Precision*	F_1
LR	2 Aug, 2018	0.5165	0.7949	0.6241
	3 Aug, 2018	0.6059	0.8094	0.6924
RF	2 Aug, 2018	0.8285	0.9695	0.8927
	3 Aug, 2018	0.8716	0.9687	0.917
DeepWalk	2 Aug, 2018	0.9114	0.8627	0.8862
	3 Aug, 2018	0.9363	0.8702	0.902
LINE	2 Aug, 2018	0.8138	0.875	0.8432
	3 Aug, 2018	0.8598	0.8761	0.8677
FdGars	2 Aug, 2018	**0.9588**	**0.9195**	**0.938**
	3 Aug, 2018	**0.9667**	**0.9135**	**0.9379**

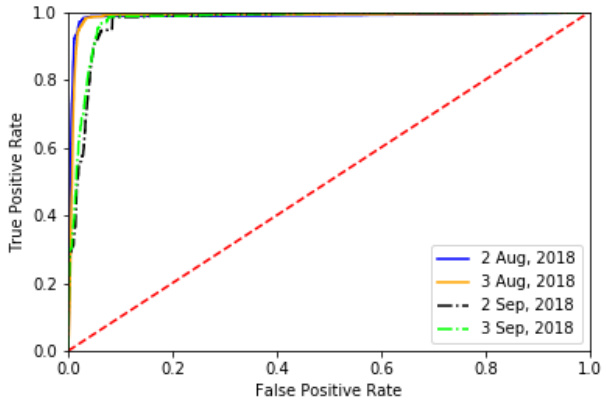


Figure 7: FdGars's ROC curves. Curves with different colors indicate the fraudster detection performance in different dates.

4 RELATED WORK

Fraudulent Users in Social Networks. Social spams became one of most popular forms due to the openness of the Internet. Fraudsters post fake/spam reviews, spread false information on blogs, forums and social media with the monetary rewards. [3, 7, 22] investigate the rumor problems, they find there exists lots of misinformation in micro-blogging and twitter platform. [3] shows that social fraudsters can even potentially alter affect the outcome of political elections.[15] analyzes the motivation of fraudsters in E-commerce (e.g. Amazon, Yelp, Alibaba), they report spammers often write fake reviews to promote their products or mislead consumers. [6] spots fraudsters in the presence of camouflage or hijacked accounts. [21] investigates content polluters, who establish links with normal users and blend the malicious information with legitimate content. Recently, fraudulent users in social networks are more adversarial, flexible and variable. A survey can be found in [10].

Graph-based Detection Methods Graph embedding methods [4, 16, 18] have been applied in lots of tasks, including node classification, link prediction, community detection, recommendation and risk control. Recently, many anomaly detection methods have focused on using graphs algorithms since it can represent and propagate the suspiciouness between objects [2]. [19] firstly captures inter-relationships among reviewers, reviews and stores based on graph model without using text information. [6] provides Fraudar for spotting fraudsters in the presence of camouflage or hijacked accounts. [23] presents NetWalK, which aims to detect structural anomalies for dynamic networks by learning network representation. [9, 17] design graph-based system (FraudEagle and FariJudge) for identifying untrustworthy users. For industrial applications, [12, 14] present graph embedding methods for detecting malicious accounts at Alipay platform.

5 CONCLUSION

In this paper, we present the first work to apply Graph Convolutional Networks on solving the problem of fraudsters detection in the online app review system. Specifically, we focus on the recall of new fraudsters. By analyzing the language styles, behaviors and relationships of reviewers, we find it is difficult to make a judgment just by utilizing single characteristic. So we propose a framework named FdGars, which combines text, behavior and relationship features of reviewers. Firstly, we extract content and behavior features for each reviewers. Secondly, we construct a graph by exploiting the relational nature of fraudsters and normal users. Thirdly, reviewers are labeled through a predefined labeling method. Based on the limited labeled reviewers, FdGards is developed to detect more fraudsters from unlabeled reviewers. Finally, we evaluate FdGars by leveraging the real-world review dataset from Tencent App Store. The results indicate that FdGars can achieve both high precision and recall. We also implement FdGars on Tencent Beacon Anti-Fraud Platform and demonstrate the effectiveness and scalability in practical scenarios. In summary, our research in this paper is expected to shed light on defending against fraudsters for large-scale online App review platforms.

ACKNOWLEDGMENTS

The review data used in this paper is courtesy of Tencent Inc. This work is supported by the National Key Research and Development Program of China (2016YFB0800102, 2016YFB0800201), the Key Research and Development Program of Zhejiang Province (2018C01088, 2018C03052), and the Major Scientific Project of Zhejiang Lab (2018FD0ZX01).

REFERENCES

[1] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. 2013. Opinion Fraud Detection in Online Reviews by Network Effects. *ICWSM* 13 (2013), 2–11.
[2] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery* 29, 3 (2015), 626–688.
[3] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. (2016).
[4] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864.
[5] Bryan Hooi, Kijung Shin, Hyun Ah Song, Alex Beutel, Neil Shah, and Christos Faloutsos. 2017. Graph-based fraud detection in the face of camouflage. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 4 (2017), 44.
[6] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. 2016. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 895–904.

[7] Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. 2013. Social Spammer Detection in Microblogging.. In *IJCAI*, Vol. 13. 2633–2639.

[8] Parisa Kaghazgaran, James Caverlee, and Anna Squicciarini. 2018. Combating Crowdsourced Review Manipulators: A Neighborhood-Based Approach. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 306–314.

[9] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahmanian. 2018. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 333–341.

[10] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* (2018).

[11] Huayi Li, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao. 2017. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1063–1072.

[12] Xiang Li, Wen Zhang, Jiuzhou Xi, and Hao Zhu. 2018. HGsuspector: Scalable Collective Fraud Detection in Heterogeneous Graphs. (2018).

[13] Yuming Lin, Tao Zhu, Xiaoling Wang, Jingwei Zhang, and Aoying Zhou. 2014. Towards online review spam detection. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 341–342.

[14] Ziqi Liu, Chaochao Chen, Xinxing Yang, Jun Zhou, Xiaolong Li, and Le Song. 2018. Heterogeneous Graph Neural Networks for Malicious Account Detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2077–2085.

[15] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S Glance. 2013. What yelp fake review filter might be doing?. In *ICWSM*. 409–418.

[16] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[17] Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*. ACM, 985–994.

[18] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[19] Guan Wang, Sihong Xie, Bing Liu, and S Yu Philip. 2011. Review graph based online store review spammer detection. In *Data mining (icdm), 2011 ieee 11th international conference on*. IEEE, 1242–1247.

[20] Guan Wang, Sihong Xie, Bing Liu, and Philip S Yu. 2012. Identify online store review spammers via social review graph. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 4 (2012), 61.

[21] Liang Wu, Xia Hu, Fred Morstatter, and Huan Liu. 2017. Detecting Camouflaged Content Polluters.. In *ICWSM*. 696–699.

[22] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on Sina Weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 13.

[23] Wenchao Yu, Wei Cheng, Charu C Aggarwal, Kai Zhang, Haifeng Chen, and Wei Wang. 2018. Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2672–2681.

Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings

Dongsheng Wang
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
wang@di.ku.dk

Qiuchi Li
Department of Information Engineering
University of Padua
Padova, Italy
qiuchili@dei.unipd.it

Lucas Chaves Lima
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
lcl@di.ku.dk

Jakob Grue Simonsen
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
simonsen@di.ku.dk

Christina Lioma
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
c.lioma@di.ku.dk

ABSTRACT

When the meaning of a phrase cannot be inferred from the individual meanings of its words (e.g., *hot dog*), that phrase is said to be *non-compositional*. Automatic compositionality detection in multi-word phrases is critical in any application of semantic processing, such as search engines [9]; failing to detect non-compositional phrases can hurt system effectiveness notably. Existing research treats phrases as either compositional or non-compositional in a deterministic manner. In this paper, we operationalize the viewpoint that compositionality is contextual rather than deterministic, i.e., that whether a phrase is compositional or non-compositional depends on its context. For example, the phrase "green card" is compositional when referring to a green colored card, whereas it is non-compositional when meaning permanent residence authorization. We address the challenge of detecting this type of contextual compositionality as follows: given a multi-word phrase, we enrich the word embedding representing its semantics with evidence about its global context (terms it often collocates with) as well as its local context (narratives where that phrase is used, which we call *usage scenarios*). We further extend this representation with information extracted from external knowledge bases. The resulting representation incorporates both localized context and more general usage of the phrase and allows to detect its compositionality in a non-deterministic and contextual way. Empirical evaluation of our model on a dataset of phrase compositionality[1], manually collected by crowdsourcing contextual compositionality assessments, shows that our model outperforms state-of-the-art baselines notably on detecting phrase compositionality.

[1]https://github.com/dswang2011/ImprovedRankedList/tree/master/input

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316584

CCS CONCEPTS

• **Information systems** → *Data encoding and canonicalization.*

KEYWORDS

Compositionality detection; Knowledge base; Word embedding

ACM Reference Format:
Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. 2019. Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316584

1 INTRODUCTION

Automatic compositionality detection refers to the automatic assessment of the extent to which the meaning of a multi-word phrase is decomposable into the meanings of its constituents words and their combination. For example, while *brown dog* is a fully compositional phrase meaning a dog of brown color, *hot dog* is a non-compositional phrase denoting a type of food. Compositionality plays a vital role in word embeddings because a non-decomposable phrase should, in principle, be treated as a single word instead of a bag of word (BOW) in word embedding approaches.

A typical line of research in automatic compositionality detection is to "perturb" the input phrase by replacing one of its constituent words at a time with its synonym, and then to measure the semantic distance between the original phrase and the perturbed phrase set [8]. The larger this distance, the less compositional the original phrase. For instance, *hot dog* would be perturbed to *warm dog* and *hot canine*. The semantic distance between the original phrase and its two perturbations is high, indicating that they denote different concepts; hence *hot dog* is non-compositional. However, the phrase *brown dog* would be perturbed to *hazel dog* and *brown canine*, which have a shorter semantic distance to *brown dog*, indicating that it is compositional.

In this paper, we posit that the compositionality of a phrase is not dichotomous or deterministic, but instead varies across scenarios. For instance, *heavy metal* could refer to a dense metal that is toxic,

which is compositional, but it could also be non-compositional when it refers to a genre of music. Previous work acknowledges this property of compositionality theoretically [8], but no operational models implementing this have been presented to this day.

Given a multi-word phrase as input, we reason that the phrase is used in some narrative, e.g., a query, sentence, snippet, document, etc. We refer to this narrative as *usage scenario* of the phrase. We combine evidence extracted from this usage scenario of the phrase with the global context (frequently co-occurring terms) of the phrase and use this to enrich the word embedding representation of the phrase. We linearly combine the weights of the tokens that are obtained from the usage scenario and the global context. We further extend this representation with information extracted from external knowledge bases.

We evaluate our model on a large dataset of phrases which are labeled as per five degrees of compositionality under various usage scenarios. We find that our model outperforms state-of-the-art baselines notably on identifying phrase compositionality. Our contributions are as follows:

- A novel model that detects phrase compositionality under different contexts and that outperforms the state of the art performance in the area.
- A benchmarking dataset of contextualized compositionality detection, that we make publicly available to the community.

2 RELATED WORK ON AUTOMATIC COMPOSITIONALITY DETECTION

Compositionality detection mainly focuses on the semantic distance or similarity calculation between a given phrase and its component words or its perturbations under a corpus or dictionary. Earlier approaches mostly estimate the similarity between the original phrase and its component words. For example, Baldwin et al. [1], and Katz and Giesbrecht [6] employ Latent Semantic Analysis (LSA) to calculate the semantic similarity (and hence to measure compositionality). Venkatapathy and Joshi [16] extended this by adding collocation features, e.g., phrase frequency, point-wise mutual information, extracted from the British National Corpus.

More recent work estimates the similarity between a phrase and perturbed versions of that phrase where the words are replaced, one at a time, by their synonyms. For instance, Kiela and Clark [7] compute the semantic distance between a phrase and its perturbation, using cosine similarity, which measures a phrase weight by pointwise-multiplication vectors of its terms. Lioma et al. [10] calculate the semantic distance with Kullback-Leibler divergence based on a language model; and, in subsequent work, Lioma et al. [8] represent the original phrase and its perturbations as ranked lists, and measure their correlation or distance.

A promising line of work uses word embeddings and deep artificial neural networks for compositionality detection. Salehi [15] employs the word-based skip-gram model for learning non-compositional phrases, treating phrases as individual tokens with vectorial composition functions. Hashimoto and Tsuruoka [5] adopt syntactic features including word index, frequency and PMI of a phrase and its components words to learn the embeddings. Yazdani et al. [17] utilize a polynomial projection function and deep artificial neural networks to learn the semantic composition and detect non-compositional phrases like those that stand out as outliers, assuming that the majority are compositional.

Closer to our work, Salehi el. [14] use Wiktionary and utilize the definition, synonyms, and translations of Wiktionary to detect non-compositional components. Specifically, they analyze the lexical overlap between the definition of a phrase and its component words to measure compositionality. They assume that multi-word phrases are included in Wiktionary, while there is no guarantee for perfect coverage of the dictionary. Unlike this approach, we use Wiktionary together with DBPedia as a structured knowledge base to represent the contextual semantics of phrases.

To our knowledge, no prior work has operationalized the compositionality of a phrase as contextual.

3 OUR CONTEXTUAL REPRESENTATION MODEL FOR COMPOSITIONALITY DETECTION

3.1 Problem Formulation

Given an input phrase p and its accompanying usage scenario s, the aim is to compute the compositionality score $Score(p)$ of phrase p with respect to usage scenario s. We follow the substitution-based line of work [7], which (a) generates perturbations of the input phrase p by substituting one word at a time with its synonym, (b) builds a semantic representation (a vector of its co-occurring terms) separately for the input phrase p and each perturbed phrase, and (c) uses the distance between the vectors of the input phrase and its perturbations to approximate the compositionality of the input phrase: the higher the distance, the less compositional the input phrase. This substitution-based line of work does not accommodate the usage scenario of the input phrase or its perturbations. The vectors of co-occurring terms are computed on one corpus, and hence these vectors represent the *global* distributional semantics of the input phrase and its perturbations. We extend this line of work by incorporating the *local* usage scenario of the phrase and its perturbations. We furthermore enrich these representations using external knowledge bases KBs. We describe this next.

Figure 1 shows how the phrase and scenario are fed into the external corpus and knowledge base in a sequential manner in the architecture, which we refer to as contextual representation model (CRM).

3.2 Building Global and Local Phrase Context

Global Phrase Context. In Natural Language Processing (NLP), the distributional semantics of an input word are computed by fixing a natural number n and, for each occurrence of a word in some corpus, finding the n words occurring immediately before, and n words occurring immediately after each occurrence of the input word (called *context window*). If there is a total of N context windows for a word, its distributional semantics in vector form can be calculated by using all these N windows. Because this is a global representation of the word's distributional semantics across the whole corpus, the vector is called a *globalized* vector. A general word embedding (e.g., word2vec) is comparable to such global context. Concretized representations of this globalized vector can be

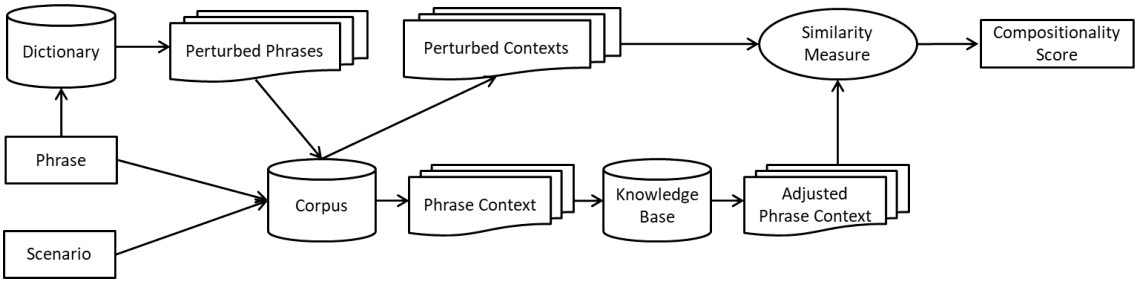


Figure 1: The diagram for the CRM sequential framework.

calculated with, e.g. ranked lists or word embeddings, as described in section 4.1.

Local Phrase Context. We aim to incorporate a representation of the local usage scenario of the input phrase (*local phrase context*) into the above described global phrase context. This representation will not be in the vector directly, because usage scenarios are typically extremely short (in terms of words), which may strongly bias the contextual representation of the phrase. Therefore, we rank all the global context windows of the phrase according to their similarity to the usage scenario of the phrase, and we select the top K most similar context windows to the usage scenario. These top K context windows are used to build the local usage scenario context representation of the phrase. Then, we linearly combine the global representation of the phrase (i.e., by taking all N context windows) and the local usage scenario representation of the phrase (i.e., the top K context windows) to acquire the *localized phrase context*. The ranking score is the similarity between the usage scenario s and a window W_i, i.e. $\text{sim}_i = \text{similarity}(W_i, s) \in [0, 1]$. The details of how the similarity score is computed are introduced in Section 4.

In the above, the value of K is determined by the length of the usage scenario as follows:

$$K = \max\left(\frac{N}{2^{\text{length}(s)}}, M\right) \tag{1}$$

where N is the total number of context windows that contain phrase p in the corpus; length(s) is the number of words in usage scenario s excluding the original phrase; and M is a threshold. We explain these next.

We posit that $2^{\text{length}(s)}$ indicates the degree of shrinking: the longer the usage scenario is, the smaller number of windows will be shrunk. The reason behind this is that the longer the usage scenario is, the more semantics it contains and subsequently fewer specific windows we are supposed to be capable of locating on. For instance, if the usage scenario is empty with length(s) = 0, then it returns the entire N windows of p with no shrinking performed; if the usage scenario has three words with $2^{length(s)} = 8$, it only collects the top $\frac{1}{8}$ of all the windows ($K = \frac{N}{8}$). Note that K depends on the usage scenario length, and is not fixed as a threshold of the similarity values. Since the similarity values may vary drastically in between $[0, 1]$ for different usage scenarios, we argue that our method is more robust to such variations. Furthermore, an empirical threshold of M is introduced to guarantee at least M windows will be selected anyhow. To this end, the localized usage scenario context of a phrase is given by:

$$C(p, s) = \alpha \frac{\sum_{i=1}^{N} \mathcal{R}(W_i)}{N} + (1 - \alpha) \frac{\sum_{j=1}^{K} \mathcal{R}(W_j)}{K} \tag{2}$$

where a is a weight parameter between 0-1 indicating the weight of global vector, and the remaining of $1 - a$ corresponds to the contribution of the localized vector; $\mathcal{R}$ denotes the semantic representation for W. In this paper, we represent a phrase as a ranked list of words and word embedding, so the same symbol $\mathcal{R}$ is adopted to denote both. The approach to calculate $\mathcal{R}$ is described in section 4.1.

3.3 Enriching Global and Local Phrase Context with Knowledge Bases

We enrich the global and local context representations of the input phrase with information extracted from external knowledge bases. We describe this next.

We reason that the corpus used to extract the global and local contexts has good coverage of various but not all possible usage scenarios of the phrase. A knowledge base is expected to contain more comprehensive, declarative information about the phrase, e.g., entities and phrase senses with categorized information. We, therefore, enrich the global and local phrase contexts extracted from the corpus with phrase information extracted from external knowledge bases.

Given an input phrase, we collect all candidate senses and entities (uniformly referred to as *candidates* in this paper) by searching the following properties (and associated values) from the knowledge bases: the properties *dbpedia:redirects*, *dbpedia:disambiguation* and their propagation relation with *dbpedia:name* and *rdfs:label*. The associated resources in these retrieved triples result in a set of candidates. Then, for each candidate, the values of *rdfs:label*, *dbpedia:abstract* and *rdf:type* are concatenated as the context for that candidate, excluding the title (which is mostly the phrase name). We also use the interface[2] to retrieve senses from Wiktionary and merge them into the same candidate set for that given input.

Most phrases only contain a limited number of candidates, and different candidates of the same phrase can have entirely different meanings or be distinct entities. We hence investigate a sequential way to incorporate the knowledge base into the phrase contextual representation, as follows. First, the phrase is fed into the knowledge base to find all candidate articles. Then, the candidates are ranked

[2]https://dkpro.github.io/dkpro-jwktl/

according to how similar they are to the localized phrase context. Those with similarity values above a certain threshold are identified as the matched candidates, denoted as $\{D_i, sim_i\}_{i=1}^{n}$, where D_i and sim_i refer to the i^{th} matched candidate with similarity value $\text{sim}_i \in [0, 1]$. A linear combination of the localized phrase context and the candidate articles is then conducted to compute the adjusted phrase context as follows:

$$C(p, D) = \lambda C(p, s) + (1 - \lambda) \sum_{i=1}^{n} w_i \mathcal{R}(D_i), \tag{3}$$

where $w_i = \frac{\text{sim}_i}{\sum_{i=1}^{n} \text{sim}_i}$ is the normalized similarity score for the i^{th} candidate article D_i while $\mathcal{R}(D_i)$ denotes the semantic representation for D_i. Since KB contains well-defined knowledge of words, we use a weighted sum of the matched candidates, instead of a simple average of matched contexts in the text corpus.

The knowledge base we employ consists of DBpedia, Wiktionary, and Wordnet. DBpedia is constructed by extracting structured information from Wikipedia. The English version of the DBpedia contains 4.58 million entries, of which 4.22 million are classified and managed under one consistent ontology. Wiktionary is a multilingual, web-based, freely available dictionary, thesaurus and phrase book, designed as the lexical companion to Wikipedia. Volunteers collaboratively construct Wiktionary, so there are no specialized qualifications necessary.

3.4 Non-linear combination

This section represents an approach of non-linear combination as a companion to the linear combination approach introduced in section 3.2 and 3.3. In addition to the weight parameter oriented design of the linear combination, we also employ a non-linear sigmoid function in RNNs (recurrent neural networks), which resolves the arrangement of the combining order for context inputs. In other words, RNNs take into consideration the feedback from the previous context vector back and forth, leading to numerous applications [3, 4]. Specifically, we train a neural network model using the Keras library to identify the compositionality label of each phrase. We encode the semantics adopting pre-trained word embeddings - word2vec [11] as word representations, a recurrent neural network with LSTM cells as the model, and cross-entropy as the loss function. As an optimizer, we utilize Adam optimizer for training the model.

In a realistic scenario (also represented in our dataset) there are fewer non-compositional than compositional phrases. This situation resembles the class imbalance issue which happens when one class (or label) is represented by most of the examples while the other one is represented just by a few. Therefore, we adopt re-sampling strategies to tackle this problem.

3.5 Compositionality Detection

Here we introduce our proposed method for compositionality detection with Algorithm 1. Given a phrase p of length l and its usage scenario s in a large corpus $Corp$, we compute its compositionality score through the following steps:

(1) Obtain localized phrase context through Eq. 1 and 2. The usage scenario of a phrase is the critical information, and the idea behind this step is to smooth the scenario context representation with the original phrase representation, as shown in line 8 in Algorithm 1.

Input: Phrase p with length l
Input: Usage scenario s for p
Input: Corpus $Corp$
Input: Knowledge Base - DBPedia
Input: Similarity threshold - $thred$
Output: Compositionality score $comp(p)$

1: Set of perturbed phrase $S(\hat{p}) \leftarrow \varnothing$
2: Find synonym $\hat{t}$ of each term $t \in p$
3: **for** each $\hat{t}$ **do**
4: Perturbed phrase $\hat{p} \leftarrow \{\hat{t}, l\text{-1 original terms } t_o\}$
5: Update perturbed phrase set $S(\hat{p}) \leftarrow S(\hat{p}) \cup \hat{p}$
6: **end for**
7: **for** phrase $p' \in \{p \cup S(\hat{p})\}$ **do**
8: $C(p') \leftarrow$ get context terms from localized phrase context from $Corp$, smooth with Eq. 1 if it has scenario
9: **end for**
10: Find n candidate articles D_i from KB where $sim_i = similarity(s, D_i) > thred$
11: $\mathcal{R}(D_i) \leftarrow$ semantic representation of D_i
12: $C(p, D) \leftarrow$ linear combined context $\lambda C(p) + (1 - \lambda) \sum_{i=1}^{n} \frac{sim_i}{\sum_{i=1}^{n} sim_i} \mathcal{R}(D_i)$
13: $Q(L_{\hat{p}}) \leftarrow \varnothing$
14: **for** each perturbed phrase $\hat{p} \in S(\hat{p})$ **do**
15: $Q(L_{\hat{p}}) \leftarrow Q(L_{\hat{p}}) \cup C(\hat{p})$
16: **end for**
17: **return** $\frac{1}{|Q(L_{\hat{p}})|} \sum_{C(\hat{p}) \in Q(L_{\hat{p}})} Similarity(C(\hat{p}), C(p, D))$

Algorithm 1: Algorithm of contextual compositionality detection

(2) Adjust phrase context with a knowledge base. The knowledge base is fed to adjust the localized phrase context where we adopt Eq. 3 to encode the information (from line 10 to line 12).
(3) Obtain a perturbed phrase set. For each term in the phrase, we find its synonyms in WordNet. We then generate the set of perturbed phrases $S(p)$ as: S(p) = {$\hat{p}$ where $\hat{p}$ = l-1 terms of p plus a synonym of the remaining term of p}, from line 3 to line 6.
(4) Construct a perturbation representation set. For each perturbed phrase $\hat{p}$ in S(p), the corresponding representation $C(\hat{p})$ is composed of all windows of $\hat{p}$ from the corpus, and is added to the perturbation list $Q(L_{\hat{p}})$ from line 14 to line 16. Note that we do not combine context from KB for perturbations.
(5) Compute the compositionality score for the input phrase, shown in line 17, using the following equation:

$$score(p) = \frac{\sum_{\hat{p} \in S(p)} sim(C(p, D), C(\hat{p}))}{|S(p)|} \tag{4}$$

4 IMPLEMENTATION

4.1 Semantic Representation

The semantic representation of a context (a phrase, a context window or a candidate content), i.e., $\mathcal{R}(\cdot)$ is concretized as either a ranked list or a word embedding. For the ranked list model, we calculate the TF-IDF as weight for all the tokens, rank them according to the weight, resulting in a ranked list of those tokens as the localized contextual representation. For the word embedding model, we use existing pre-trained word vectors - Glove [12], and represent the vector with the average of all tokens. The corpus we employ in our experiment is ClueWeb12-B13, a subset of some 50 million pages of ClueWeb12-Full dataset[3].

These two contextual representations lead to two different compositionality scores for the same model. We apply suffixes "-word embedding" or "-ranked list" in order to distinguish the way the contextual representation is computed, resulting in two distinct models, namely *CRM word embedding* and *CRM ranked list*.

4.2 Similarity Measure

In this study, we are faced with the problem of computing the similarity value between two context vectors. Here, we consider two types of similarity measures to achieve this purpose: *cosine similarity* and *Pearson correlation coefficient*.

One of the most commonly used similarity measures, *cosine similarity*, computes the cosine value of the angle between the two vectors of the same length. For two vectors $\vec{a} = [a_1, a_2, .., a_n]$ and $\vec{b} = [b_1, b_2, .., b_n]$, their cosine similarity cossim(a,b) is given below:

$$cossim(a, b) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{5}$$

The *Pearson correlation coefficient* computes the degree of correlation between two variables, each having a set of observed values. Suppose two variables X and Y are associated with two set of values $\{X_1, X_2, ..., X_n\}$ and $\{Y_1, Y_2, ..., Y_n\}$ respectively. The Pearson correlation coefficient r can, therefore, be computed as follows:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{6}$$

where $\overline{X}$ and $\overline{Y}$ denote the average of X and Y respectively.

4.3 Perturbation

Here, we introduce the process to obtain the perturbations of a phrase p with length l. First, we get the synonyms for each word in the phrase. Then, we construct the whole perturbation set, which contains all phrases composed of $l-1$ words in p and a synonym of the remaining word. Suppose the i^{th} word has n_i synonyms, then the perturbation set contains $\sum_{i=1}^{l} n_i$ perturbed phrases. We then prune the perturbation set by filtering out the rare perturbed phrases in the text corpus. Basically, we compute the occurrence frequency of all perturbed phrases and pick the perturbed phrases with top K frequency values. In our study, we set K to be 7, which is derived from empirical observation of the data. Then, the final perturbation set contains 7 perturbed phrases in our study.

[3]https://lemurproject.org/clueweb12/

4.4 Parameter Settings

Contextual Windows setting: We set $window = 20$, which means it scans the previous 20 and subsequent 20 words of that phrase, with a sum of 40 words for each window. In Equation 1, we set $M = 10$, and the base 2 can also be parameter-free which can be changed into 2,3,4,etc., to increase the localization level.

Knowledge base threshold: As for the threshold of KB candidates, we set the threshold of similarity value between localized context and KB candidates as 0.5 to filter out those candidates with similarity less than 0.5.

Ranked list length: In line with the work [8], we set a maximum length of the ranked list as 1000, which means that we rank the tokens according to their TFIDF weight, and the tokens after the position of 1000 would be pruned.

Table 1: Results of different compositionality detection methods; *na* denotes *not applicable*.

Unsupervised Methods	ρ	α, λ
Baseline: Ranked list [8]	0.131	na
Baseline: Word Embedding	0.147	na
CRM ranked list	0.209	0.1,0.5
CRM Word Embedding	0.375	0.9, 0.1
Supervised Methods (20% Testing)	ρ	
RNN (LSTM cells)	0.176	na
RNN (LSTM cells) CRM	0.324	na

Training and testing: As we are working with imbalanced data, we use a random oversampling strategy. We split our data in a stratified fashion into 65% for training, 15% for validation, and 20% for testing. The re-sampling is be done after splitting the data into training and test, and only on the training data, i.e., none of the information in the test data is being used to create synthetic observations.

5 EXPERIMENT AND VALIDATION

In this section, we evaluate the effectiveness of the model presented in Section 3. Section 5.1 introduces the dataset and Section 5.2 presents the results achieved by our model.

5.1 Crowdsourcing data

We employ a dataset that consists of 1042 phrases that are noun-noun 2-term phrases [2]. In this dataset, each phrase was assessed four times using a binary scale (compositional or non-compositional). However, these phrases are assessed with a deterministic label, meaning that no scenario or context was given, and the degree of compositionality may not always be binary [13]. Therefore, we extend the dataset into a new version where each phrase is enriched with one or two scenarios if possible, by taking advantage of a

Table 2: Summary of dataset statistics.

No. Non-Compositional	43 (3.6%)
No. Mostly Non-Compositional	145 (12.1%)
No. Ambiguos Phrases	126 (10.5%)
No. Mostly Compositional	141 (12.0%)
No. Compositional	739 (61.8%)
Unique number of Phrases	1042
No. of context	1194
Average number of context by Phrase	1.146

crowdsourcing website - Figure Eight [4], and we use a graded level of compositionality. In Table 2 we summarize the dataset statistics.

We divided the assessment into two stages: for the first stage, the trustful assessors, with level 3 (highest in Figure Eight), are required to understand the various meanings of a phrase, and, if possible, create two scenarios for the same phrase. From these two scenarios, one should be compositional or as compositional as possible, and the other non-compositional or as non-compositional as possible. If the phrase can only be compositional or only be non-compositional, then they create one scenario for it. For the second stage, the assessors are required to assess the compositionality of phrases within different scenarios with one of the five graded labels: compositional, mostly-compositional, ambiguous to judge, mostly non-compositional, and non-compositional. Note that, for the first stage, the two scenarios of a phrase are not necessarily of two extreme polarities.

[4]https://www.figure-eight.com/

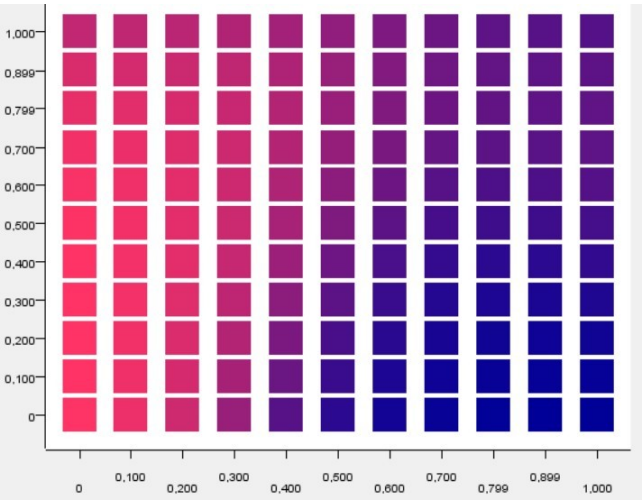


Figure 2: The grid search for Word embedding based Contextual representation. x-coordinate is α for controlling localized context and λ stands for y-coordinate, controlling the KB combining weight. Deeper blue represents higher performance whereas red indicates the opposite.

5.2 Performance and Validation

Two linear combination parameters influence the performance of our model: the combination weight α (in Eq. 2) between the vectors of a phrase and its scenario, resulting in a localized phrase context, and λ (in Eq. 3) between the localized context and knowledge base. The impacts of these two parameters on the final performance are visualized in Figure 2 and 3, corresponding to the word embedding-based and ranked list-based contextual representation respectively. α and λ denote the x and y coordinates. The colors indicate the performance, which is the correlation between the ground truth labels and the predicted labels of our models ranging from -1 to 1. The performance values are colored ranging from red to blue, representing the lowest performance to the highest.

As shown in Figure 2, the performance is negatively correlated with α while positively correlated with λ. This indicates that reducing the relative importance of localized context (right direction on x-coordinate) while enhancing the influence of knowledge base (bottom direction on y-coordinate) can improve the performance for word embedding based contextual representation. In contrast, as shown in Figure 3, if we ignore the *0* column, α is negatively correlated with the performance while λ does not have an apparent influence on the performance. This indicates that attaching higher importance to the localized context (left direction on x-coordinate) can improve the performance for the ranked list based contextual representation, while the adoption of knowledge base does not have an apparent influence to the overall performance. The first *0* column, which is shown more like an outlier, indicates that the existence of a vector of the original phrase is necessary. In other words, localized context would have relatively poor performance.

As summarized in Table 1, the performance improved from 0.147 to 0.375 for CRMs based on word embeddings (the best); from 0.131 to 0.209 for CRMs based on ranked list; and 0.176 to 0.324 for CRM based on RNN. For the word embedding based contextual representation model, relying more on the knowledge base while keeping the scenario to limited importance will lead to a high-performed model; for the ranked list based contextual representation model, on the other hand, adequately high adoption of localized context can lead to improved performance. The reason behind this can be that the knowledge base contains relatively trimmed but well-categorized information, therefore, the word embedding model can take full use of this text as informative vectors. In contrast, ranked lists, depending on tokens, work better on a large-scale corpus where they induce a large number of context windows. However, the knowledge base contains a limited number of tokens that may have little contribution to the final representation. Even though we can tune the weight of tokens from a knowledge base, it still can have limited influence in comparison to the long ranked list, which can be as long as 1000 tokens in our experiment.

For the non-linear combination where we employed the sigmoid function in RNNs, the CRM based on RNN still beats the original RNN. However, the performance is still lower than the unsupervised approaches.

6 CONCLUSION

We developed a novel method for compositionality detection where the compositionality of a phrase is contextual rather than static.

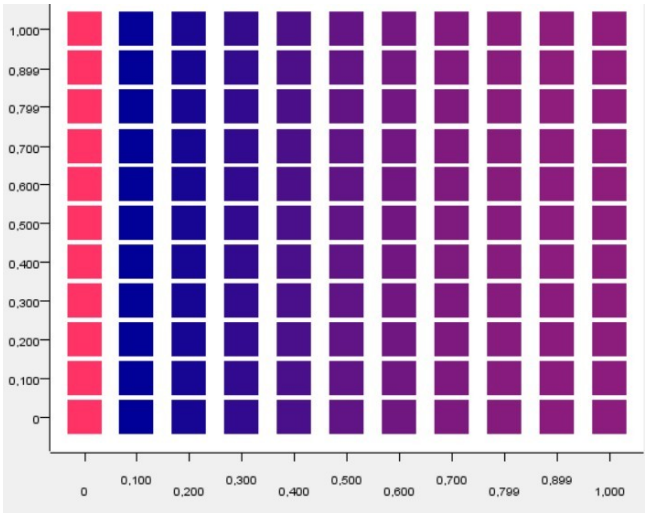


Figure 3: The grid search for ranked list based Contextual representation. x-coordinate is α for controlling localized context and y-coordinate stands for λ, controlling the KB combining weight. Deeper blue represents higher performance whereas red indicates the opposite.

Instead of considering an isolated phrase as input, we assume a phrase and its usage scenario (e.g., a query, snippet, sentence, etc.) as input, and we model a joint semantic representation of these by combining distributional semantics extracted from a corpus and additional evidence extracted from an external structured knowledge base.

Our resulting model uses word embeddings to detect compositionality, more accurately than the related state of the art. Our experiments show that for word embeddings, the usage of knowledge bases can lead to notable performance improvements.

In the future, we plan to evaluate our model on further datasets and compositionality detection scenario, e.g., Verbal Phraseological Units (VPUs).

ACKNOWLEDGEMENT

This work is supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 721321.

REFERENCES

[1] Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*. Association for Computational Linguistics, 89–96.

[2] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A Multiword Expression Dataset: Annotating Non-Compositionality and Conventionalization for English Noun Compounds. In *Proceedings of the 11th Workshop on Multiword Expressions (MWE-NAACL 2015)*. Association for Computational Linguistics.

[3] Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. [n. d.]. Neural Speed Reading with Structural-Jump-LSTM. In *Proceedings of the 2019 International Conference on Learning Representations, ICLR 2019*.

[4] Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Modelling Sequential Music Track Skips using a Multi-RNN Approach. In *Proceedings of the 2019 International Conference on Web Search and Data Mining, WSDM 2019, Sequential Skip Prediction Challenge*. in press.

[5] Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. *arXiv preprint arXiv:1603.06067* (2016).

[6] Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, 12–19.

[7] Douwe Kiela and Stephen Clark. 2013. Detecting Compositionality of Multi-Word Expressions using Nearest Neighbours in Vector Space Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1427–1432. http://aclweb.org/anthology/D13-1147

[8] Christina Lioma and Niels Dalum Hansen. 2017. A Study of Metrics of Distance and Correlation Between Ranked Lists for Compositionality Detection. *Cogn. Syst. Res.* 44, C (Aug. 2017), 40–49. https://doi.org/10.1016/j.cogsys.2017.03.001

[9] Christina Lioma, Birger Larsen, and Peter Ingwersen. 2018. To Phrase or Not to Phrase – Impact of User versus System Term Dependence Upon Retrieval. *Data and Information Management* 2, 1 (2018), 1–14.

[10] Christina Lioma, Jakob Grue Simonsen, Birger Larsen, and Niels Dalum Hansen. 2015. Non-compositional term dependence for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 595–604.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[13] Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*. Chiang Mai, Thailand. http://aclweb.org/anthology-new/I/I11/I11-1024.pdf

[14] Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1792–1797.

[15] Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 977–983.

[16] Sriram Venkatapathy and Aravind K Joshi. 2005. Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 899–906.

[17] Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1733–1742.

First International Workshop on e-Commerce and NLP (ECNLP) Chairs' Welcome

It is our great pleasure to welcome you to the First International Workshop on e-Commerce and NLP (ECNLP).

This workshop focuses on intersection of Natural Language Processing (NLP) and e-Commerce. NLP and information retrieval (IR) have been powering e-Commerce applications since the early days of the fields. Today, NLP and IR already play a significant role in e-Commerce tasks, including product search, recommender systems, product question answering, machine translation, sentiment analysis, product description and review summarization, and customer review processing. With the exploding popularity of chatbots and shopping assistants – both text- and voice-based – NLP, IR, question answering, and dialogue systems research is poised to transform e-Commerce once again.

The ECNLP workshop was designed to provide a venue for the dissemination of late-breaking research results and ideas related to e-Commerce and online shopping, as well as a forum where new and unfinished ideas could be discussed. It has brought together researchers from both academia and industry.

We have received a larger number of submissions than we could accept for presentation. ECNLP 2019 received 23 submissions of long and short research papers and of position papers. In total, ECNLP 2019 featured 14 accepted oral and poster presentations (61% acceptance rate).

The selection process was competitive and we believe it resulted in a balanced and varied program that is appealing to audiences from the various sub-areas of e-Commerce. We would like to thank everyone who submitted a paper to the workshop. We would also like to express our gratitude to the members of the Program Committee for their timely reviews, and for supporting the tight schedule by providing reviews at short notice.

We hope that you enjoy the workshop!

Shervin Malmasi
Amazon Research

Nicola Ueffing
eBay Inc.

Oleg Rokhlenko
Amazon Research

Eugene Agichtein
Emory University

Ido Guy
eBay Inc.

THE WEB CONFERENCE

ECNLP 2019 Organization

Workshop Chairs:

Shervin Malmasi (Amazon Research)
Nicola Ueffing (eBay Inc)
Oleg Rokhlenko (Amazon Research)
Eugene Agichtein (Emory University)
Ido Guy (eBay Inc)

Program Committee:

Anuj Kumar Goyal (Amazon Alexa)
Bei Shi (Tencent AI Lab)
David Martins de Matos (Instituto Superior Técnico, Universidad de Lisboa)
Els Lefever (LT3, Ghent University)
Irina Temnikova (Sofia University, Bulgaria)
Julia Kreutzer (Heidelberg University)
Kevin Small (Amazon)
Marzieh Fadaee (University of Amsterdam)
Raksha Sharma (IIT Bombay)
Rongting Zhang (Amazon)
Selcuk Kopru (eBay)
Sheila Castilho (The ADAPT Centre/ Dublin City University)
Shuo Zhang (Bose Corporation)
Wai Lam (The Chinese University of Hong Kong)
Yanbin Lu (Amazon)

On the Relation Between Products' Relevance and Customers' Satisfaction in Voice Shopping

ECNLP'19 Workshop

David Carmel
Amazon Research
Haifa, Israel
dacarmel@amazon.com

ABSTRACT

In this talk we consider the domain of voice shopping in Alexa, Amazon's voice assistant. In this domain, search scenarios are integral part of the shopping sessions, where users seek for a product to buy, or for some information about a product. The fact that in voice search, both the input and output are spoken, involves many challenges in automatic speech recognition, natural language understanding, question answering, and new user experiences. We will elaborate on customers' behavior in voice shopping, where we have observed an interesting and surprising phenomenon that many customers purchase or engage with irrelevant search results. The term "irrelevance" may mislead, since a relevant item is typically interpreted as "anything that satisfies the user needs". Thus, the title of this work may look as an oxymoron – the purchase of a product is a strong signal of relevance to the customer. In the context of this work we take a simplified approach. We mark product items as relevant or irrelevant to the user query based on the relevance judgments of several human annotators. However, even in the context of objective relevance judgments, it is still surprising that so many customers engage with irrelevant results. We will analyze this phenomenon and demonstrate its significance. We will offer several hypotheses as to the reasons behind customers' purchase and engagement with irrelevant results, including customers' personal preferences, trendiness of the products and their relatedness to the query, the query intent and the product price.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; *Information retrieval*;

KEYWORDS

Voice shopping, user satisfaction, product relevance

ACM Reference Format:
David Carmel. 2019. On the Relation Between Products' Relevance and Customers' Satisfaction in Voice Shopping: ECNLP'19 Workshop. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316610

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316610

1 BIOGRAPHY

David is a Principal Applied Scientist at Amazon Research, and an ACM Distinguished Engineer. David's research is focused on search, question answering, query performance prediction, and text mining. For several years David taught the Introduction to IR course at the CS department at Haifa University. David has published more than 100 papers in IR and Web journals and conferences, and serves on the editorial board of the IR journal and as a senior PC member or as Area Chair of many ACM conferences (SIGIR, WWW, WSDM. CIKM). He organized a number of workshops and taught several tutorials at SIGIR, and WWW. David is co-author of the book "Estimating the Query Difficulty for Information Retrieval", published in 2010. David earned his PhD in Computer Science from the Technion, Israel Institute of Technology in 1997.

Shop your Right Size: A System for Recommending Sizes for Fashion products

G. Mohammed Abdulla*
gmohamedabdulla@gmail.com

Shreya Singh
Myntra Designs, India
shreya.singh1@myntra.com

Sumit Borar
Myntra Designs, India
sumit.borar@myntra.com

ABSTRACT

Size selection is a critical step while purchasing fashion products. Unlike offline, in online fashion shopping, customers don't have the luxury of trying a product and have to rely on the product images and size charts to select a product that fits well. As a result of this gap, online shopping yields a large percentage of returns due to size and fit. Hence providing size recommendation for customers enhances their buying experience and also reduces operational costs incurred during exchanges and returns. In this paper, we present a robust personalized size recommendation system which predicts the most appropriate size for users based on their order history and product data. We embed both users and products in a size and fit space using skip-gram based Word2Vec model and employ GBM classifier to predict the fit likelihood. We describe the architecture of the system and challenges we encountered while developing it. Further we also analyze the performance of our system through extensive offline and online testing, compare our technique with another state-of-art technique and share our findings.

KEYWORDS

Personalization, Size Prediction, Word2Vec, Doc2Vec

ACM Reference Format:
G. Mohammed Abdulla, Shreya Singh, and Sumit Borar. 2019. Shop your Right Size: A System for Recommending Sizes for Fashion products. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316599

1 INTRODUCTION

The online fashion industry is growing at an astounding pace. Reliable statistics [11] show that the worldwide revenue is expected to rise from $481.2 billion in 2018 to $712.9 billion by 2022. Fashion products like apparels and shoes are available in various sizes and size selection becomes a critical part while purchasing them. While purchasing offline, consumers typically try the product before making a purchase whereas during online shopping, users have to rely on size charts or images of products to make the size decision. Size charts require customers to remember their body measurements and compare them with product dimensions. Moreover the fashion industry lacks standardization in terms of sizing [5] and the attributes associated with fashion products are highly subjective. Moreover, even for the same brand, different product lines and various fits (Slim, Regular etc) make choosing size a tricky process. To make the online purchase convenient, online platforms typically provide free returns. This results in higher return rates due to size and fit issues in online e-commerce and adds significant operational costs. In this paper, we propose a novel approach to recommend sizes using latent features and gradient boosting classifier(GBC). Figure 1 illustrates the size recommendation displayed on Myntra's mobile app. Every product in the catalog has multiple sizes and a combination of the product and its size is referred as SKU. The GBM classifier takes as input a combination of user and SKU vectors and returns the probability of fit. Further, the SKU vector is a combination of observable features like size, occasion, fit , material etc. and latent features learnt from purchase history using Word2Vec. The latent features transform the users and products into a size and fit space. This space has an inherent property of bringing similar-sized products closer to each other. Moreover, users whose size preferences are similar are also closer in this space. Our experiments demonstrate that models built with both observable and latent features perform better than models built solely on either observable or latent features.

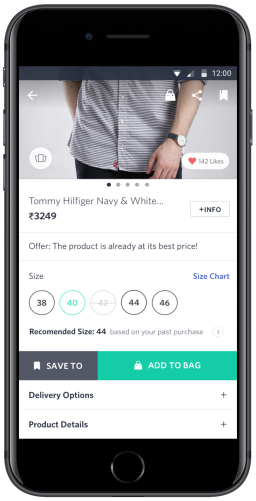


Figure 1: Size Recommendation on our platform.

Myntra is India's largest fashion e-commerce portal serving millions of customers daily. We operate solely in the Indian retail

*Work done while author was at Myntra

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316599

market and currently have support for English language on our platform. In order to provide a better shopping experience it is imperative that we help our customers choose the right size for any given product. In our journey to develop this system, we came across several challenges specific to our domain:

- Users mostly tend to purchase from similar price ranges or similar brands. Hence, Word2Vec fails to capture similarity across dissimilar price ranges or brands, resulting in local clusters in the size and fit space.
- Some articles like Shirts, T-shirts, Dresses are bought very often while others like Jackets, Sweatshirts are not purchased frequently. This poses a challenge to recommend sizes with high confidence in less often purchased categories.
- Users in our platform do not just buy for themselves, but use the same account to buy for many people. It is not trivial to make size recommendations for such users.

This paper is organized as follows. Section 3 outlines the size recommendation algorithm as a classification problem. We describe various challenges specific to our domain and our solutions in Section 4. We present a comprehensive architecture of the online size recommendation system in Section 5 followed by results and analysis in Section 6.

2 RELATED WORK

There is substantial research on personalized product recommendation for users in fashion e-commerce. Recommendations based on users' past interactions and affinities have been studied extensively [1, 2, 7]. All these recommendation systems try to model user product preferences and not size preferences. A recommendation system that recommends product sizes like {Small, Fit, Large} to customers in Amazon is presented in [12]. In this work, the authors have defined Hinge Loss and Logistic loss variants to compute true sizes for customer-product pairs by minimizing the loss function which is based on the difference between the true sizes of the customers (derived from order history) and the true sizes of the products (derived from the product dimensions). This approach is based only on size measurements and does not take into account attributes like fit, material, brand while making recommendations. Apart from these, in the textile and garments industry, several methods to find the appropriate size and fit preferences are proposed which rely on the 3D modelling of body shapes [4]. These approaches are based heavily on deriving body shapes from a database of body shape metrics which are curated manually or deriving body shapes from images. There have also been attempts to model user size preferences in industry by the likes of True Fit and Fit Analytics. However, their approach requires users to provide body measurements explicitly via surveys or questionnaires.

3 METHODOLOGY

Our method to solve the size recommendation problem is to model it as a binary classification problem with positive class consisting of SKUs retained by the users and negative class consisting of SKUs returned/exchanged by the users due to size and fit issues. Post training, we use Gradient Boosting model [8] to determine the fit probabilities. The user-SKU pair which gives us the highest confidence is the recommended size. The approach could be divided into three main components: generation of observable feature vectors, generation of latent feature vectors, and using the ensemble method for classification. We have used implicit matrix factorization as the baseline approach.

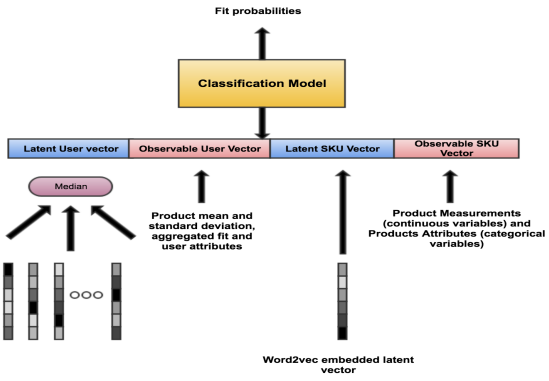


Figure 2: Classification model architecture showing the training vector as a combination of User and SKU vectors

3.1 Generation of Observable feature vector

The observable feature vector for a SKU consists of what we "observe" in any product. For our use case, the observable features consist of measurement attributes(like chest, bust, waist dimensions), material and colour. Measurement attributes are continuous values and are not uniform across different categories. For instance, in Men Topwear category, the measurements are chest, shoulder and length; whereas in Women Topwear category it is bust, shoulder and length. Other product attributes are categorical values and are consistent across categories. A combination of these gives the SKU observable features. The user observable feature vector is formed by aggregating all the observable feature vectors of the SKUs bought by a user. To this, we append additional information like mean and standard deviation of the measurement attributes which will capture the user's purchased size distribution.

3.2 Generation of Latent feature vector

Latent feature vectors incorporate the "hidden" representations of user purchase traits. We use skip-gram based Word2Vec [9] model to derive the latent feature vectors. The input to these models are the set of SKUs bought by the user. The SKUs which have been returned or exchanged are removed while generating latent feature vectors. They are however used as negative samples while training the model. These SKUs are not used directly, but are first converted to a combination of Brand, Size, Fit and Usage attribute for that product. For example a SKU with brand "Nike", size "L", fit "Slim" and usage "Sports" can be converted into a word like "Nike-Sports-Slim-L". We have analyzed that a word consisting of all these attributes for a SKU representation gives us better results than only including size and brand or size and occasion. These are necessary features as there are several product lines and usage attributes in a brand which are pertinent to a product. Hence, training a Word2Vec model with these features would help us govern the size profile of a user. The SKUs can be considered as words and all the SKUs

purchased by a user can be considered as a document. With these words and documents as inputs we train the Word2Vec network, the final activation layer of every SKU is obtained as an "off-the shelf" feature which will serve as the latent feature vector.

Formally, the user purchase data can be represented as a sparse matrix W where each row represents a user and each column represents the SKUs bought by the user. We replace each SKU by the word as explained above. Each row i of matrix W would be a document of word representations of the SKUs bought by the user. We train the Word2Vec network with matrix W where W_i represents all the purchases of a user i sorted by date which can be represented by the sequence $W_{i1}, W_{i2}, W_{i3}, ..W_{in}$, here each $W_{ij} \in C$ where C is the entire platform catalog for a particular article type and W_{ij} is the word representation of the SKU. The objective of the skip gram model is to maximize the log probability.

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{-q\leq k\leq q, k\neq 0}\log p(w_{i,j} \mid w_{i,j+k}) \tag{1}$$

where q is the hyperparameter denoting length of the purchase window. Larger q results in SKUs spanning over wide range of purchases to be considered as having same size and fit. The formulation of $p(w_j \mid w_{j+k})$ is given using Softmax function:

$$p(w_j \mid w_{j+k}) = \frac{e^{u_j, v_{j+k}}}{\sum_{k\in W} e^{u_k, v_k}} \tag{2}$$

where u and v are the input and output one hot encoded vector representation of w. After we have trained the neural network, we compute the activation of the hidden layer for each SKU $c_p \in C$ and form a latent feature vector representation f_p. Post training, the activation layer is scraped off and used as the latent vector representation for each SKU. The latent user vector can be computed by taking the mean of the latent vectors of SKUs bought by the user.

3.3 Recommendation as Gradient Boosting classification

We model the size recommendation system as a binary classification problem. For this purpose; we use Gradient Boosting Classifier for training as it will provide non-linearity and reduce over fitting. In our binary classification setting, positive class for recommendation is the set of user-SKU pairs purchased and retained by users and negative class is the set of user-SKU pairs returned or exchanged by users. When any product is returned or exchanged on our platform; the user is required to select a reason from a pre-defined set of reasons like Size/Fit issues, wrong product, late delivery etc. All the return and exchange reasons are logged and we make use of these reasons to generate negative class samples. Post training, we rank the fit probabilities of all SKU-user pairs for a given user and product (having multiple SKUs) and deem the SKU with highest fit probability as the recommended SKU(size). In our GBC model, the training vector is a concatenation of user and SKU vectors. Figure 2 shows our baseline architecture and outlines the structure of the training vector as a combination of user and SKU vectors. Further, the user and SKU vectors are themselves built of latent and observable parts as illustrated in Figure 2. In our current system, we require one SKU bought and retained by the user to make reliable recommendations for him/her.

3.4 Implicit Matrix Factorization

To compare our model we have used Implicit matrix factorization [6]. The data can be represented as a matrix with rows as users, columns as products and each cell in this matrix is the number of times given user has purchased the given product. As directly using every product on our platform resulted in a sparse matrix with sparsity greater than 99.9%; products were replaced with words as described above (example: "Nike-Sports-Slim-L"). This resulted in sparsity less than 99.9% and this sparsity is considered as an empirical standard in collaborative filtering problems. Validation data was used to tune various parameters of the matrix factorization model. The result of this method is compared with our classification approach in Section 5.6.

4 CHALLENGES

In this section we address various challenges we faced while building the size recommendation algorithm. Quantitatively we want to improve two metrics - precision and coverage. Precision in our context refers to how accurately we can make size predictions and Coverage is the number of users for whom recommendations can be made with said precision. The following challenges and their corresponding solutions improve either one or both of the performance metrics.

4.1 Local Clustering in Embedding Space

On our platform customers can be largely segmented into mass, mass premium, premium, bridge to luxury and luxury segments. These segments primarily indicate the price bands and brands which certain group of users would be interested in. For instance, in case of Jeans, brands like GAS, Replay, Scotch and Soda, where an average product is priced at 10000 INR, are considered to be luxury brands, while brands like Levis, Roadster, Wrangler and Tommy which are priced around 3500 INR are segmented into mass premium brands. There has been previous work [3] which show that users tend to purchase in their respective segments and rarely move across them while making future purchases. This has an unfortunate effect of less co-purchase history across brands belonging to different segments. Figure 3 illustrates this behaviour by aggregating cosine similarity metric between brands belonging to different segments. Techniques like Word2Vec and Collaborative filtering fail in such cases where user interactions are sparse which is evident in Row 1 of Figure 3. Such behaviour occurs due to the formation of local clusters in product embeddings.

For illustration, in Figure 3, we have considered embedding vectors generated using the method mentioned in above section for four different brands of jeans - Roadster, Wrangler, Replay and GAS. Among these brands Roadster and Wrangler fall into mass premium segment while Replay and GAS are in luxury segment. The first row shows cosine similarity between embeddings for various sizes of these brands. It can be clearly seen that brands from same segment have linear relation between various sizes, but when we consider brands across segments - the similarity values are random and represent misinformation in the embedding space. Further,

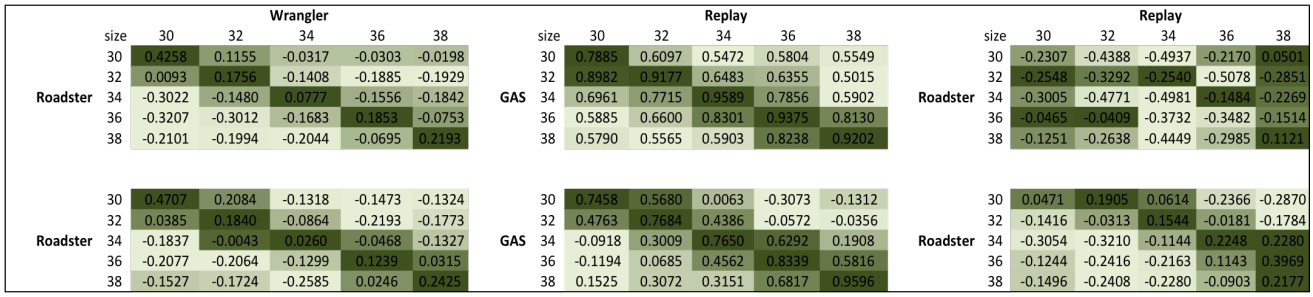


Figure 3: Cosine similarity of embedding vectors across different brand-size combinations. Row 1 shows the similarity values in normal Word2Vec method and row 2 shows the similarity values of embeddings which are generated from sampled data.

we aggregate these product embeddings to form error-prone user embeddings. This is an undesired behaviour and results in wrong recommendations. To tackle this problem we have come up with two solutions - improving the aggregation logic of user vectors and improving the underlying product vector generation logic.

4.1.1 Improve User Vector Aggregation. In the current formulation where product vectors form local clusters, there exist three different scenarios while making the final size recommendation. First scenario occurs when a user has purchased from a subset of similar brands and the product for which recommendation has to be made is a brand similar to user's past purchases. Second scenario occurs when a user's past purchases are from a subset of similar brands but the product for which recommendation has to be made is not similar to user's past purchases. Third scenario occurs when a user's past purchases are from various segments (in this case the user embedding will be incorrect) and the product to be recommended for could be any brand in the catalog. The first and second scenarios will not result in wrong recommendation because the final Gradient boosted classifier will learn to recommend correctly in these situations, but, in the third scenario, the mean of segmented vectors falls in a random location in the size and fit space resulting in incorrect recommendations.

To make our model robust for the third scenario, instead of taking mean as an aggregation metric, we experimented with other aggregation techniques. While median is less susceptible to outliers than mean, in our case taking a median of embeddings is equivalent to choosing a dominant subset which results in loss of valuable information. We also experimented by concatenating all the product vectors purchased by the user to generate a user embedding. This does not lose out on information but adds complexity to the model and also makes it sensitive to the order in which products are concatenated. Finally, we generated user embedding along with product embedding using Doc2vec[10]. In our setting, documents are all the purchases made by a user and words are SKUs represented by their attributes. Unlike Word2Vec where we only learn representation for words(SKUs), in Doc2Vec we also learn representation for the document(user in our case). In the Results section, we compare the three aggregation logic and illustrate that Doc2vec vectors are the best.

4.1.2 Improve Product Vector. Our other approach to solve the local clustering problem is to improve the underlying product vector generation logic. As co-purchase history is sparse, we came up with a method to sample from the given purchase data resulting in global clusters. The procedure for sampling is given in Algorithm 1. We create a co-purchase Graph $\mathcal{G}$ with nodes as brands and edge weights as number of times these two brands were purchased together. We also create a map $\mathcal{P}$, which has key as pair of brands and values as a list of all the users who have purchased these two brands. The sampling algorithm looks at every brand(node) in the graph and computes a median edge weight from all the outgoing edges from the current brand. It then looks at each edge, if the weight of outgoing edge is less than the median, it over samples users from $\mathcal{P}$. The output of the algorithm is a list of users $\mathcal{U}$ and for each user in this list we replicate their previous orders. This results in Word2Vec learning relation between brands which were not obvious earlier. Row 2 of Figure 3 illustrates cosine similarity computed using embeddings learned from combination of purchase history and sampled data $\mathcal{U}$.

Algorithm 1 Sampling Algorithm

Require: Graph $\mathcal{G}\{V, E\}$, Dictionary $\mathcal{P}$
  Initialization $\mathcal{U} = list$ {initialize empty list}
  for all $a \in \mathcal{G}.V$ **do**
    $m = median(\mathcal{G}, a)$ {compute median value of weights from node a}
    for all $b \in \mathcal{G}.V$ **do**
      {G(a,b) - return weight between node a and b}
      if $a \neq b$ & $\mathcal{G}(a, b) \leq m$ **then**
        v = sample $(m - \mathcal{G}(a, b))$ users from $\mathcal{P}(a, b)$
        $\mathcal{U}v$
      end if
    end for
  end for

4.2 Lack of Purchase History

In Section 3 we mentioned that we use purchase history of an article (like Shirts, T-shirts, Dresses) to create Word2Vec embeddings of products. These embeddings are feasible in articles which are sold more often in the platform.

However, many articles like Sweaters and Sweatshirts are purchased less frequently and we do not have enough co-purchase history to create quality embeddings. Hence, to enrich embeddings

for these articles, we use information of other similar articles. All the articles in our catalog belong to certain high-level categories: namely Men's Topwear , Women's Topwear, Men's Bottomwear, Women's Bottomwear and Footwear, and we train a Word2Vec model on category purchase history rather than article purchase history resulting in rich embeddings for products in all articles. Although using category for learning product embedddings is superior in theory, implementation of this method has a few challenges. Users can make purchase for multiple people using the same account or may make purchase for different articles for different people. We want to identify a dominant persona from all the different personas that exist for an account. Identifying dominant persona can be formulated as an optimization problem as below.

Let A be the set of all orders of a user then dominant user set is given by

$$\begin{aligned} & |A'|; A' \in \mathcal{P}(A) \\ \text{s.t.} \quad & abs(A_i - A_j) \leq \lambda \\ & \forall A_i, A_j \in A', i \neq j \end{aligned} \tag{3}$$

where λ is the threshold set for optimization and $\mathcal{P}(A)$ is the power set of A.

A is the set of sizes purchased by a user in a given category. For Men's Topwear, set A would contain chest measurements in inches of all the purchased products. The threshold λ is a tuneable parameter and is set with domain knowledge. Identification of dominant profile has to run for millions of users and we present a pseudo code Algorithm 2 which runs in log-linear time complexity. Once the dominant profile and its orders are identified, these orders are used in creating Word2Vec embeddings.

Algorithm 2 Optimization Algorithm

Require: List A, threshold λ

Initialization $i = 0$,$j = 1$, $maxSet = list$,$curSet = list$ {initialize empty lists}
$A = sort(A)$
while $i < len(A) - 1$ & $j < len(A)$ **do**

  if $A_j - A_i \leq \lambda$ **then**
    $curSetA_j$
    $j1$
  else
    $curSetA_i$
    $i1$
  end if
  if $len(maxset) < len(curSet)$ **then**
    $maxSet = curSet$
  end if
end while

4.3 Short Lifetime of Users and Multi-persona

About 65% of users on our platform purchase for more than one person using the same account. The current formulation of size recommendation can only make recommendations for the remaining 35% of the users. To make recommendations for multi-persona users we have developed a sub-system where users can create multiple profiles in the app. The user can tag the bought products to these profiles post placing an order. We aim to target multi-persona users through this approach. Each of these profiles can now be considered as a single-persona user to get profile-level size recommendation. It is observed that the average lifetime of users on our platform is short. We want to make size recommendation available to users during early stage after adoption and we leverage profile-level size recommendation to achieve this task. In the later section we explain in detail how profile-level recommendation system is implemented.

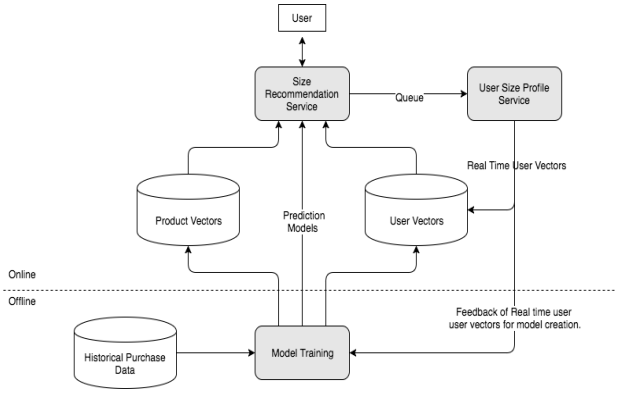


Figure 4: Simplified System Architecture

5 SYSTEM ARCHITECTURE

In this section we outline the components of size recommendation system using the the block diagram in Figure 4. The system consists of an offline training system and an online serving system.

In the offline training system, we use historical purchase data and content data of products to create embeddings. These embedddings are then used by the gradient boosted classifier model to create the prediction models. The prediction models, user vectors and product vectors are fed into databases for the online serving systems.

In the online system, size recommendation service is deployed across multiple servers coupled with a load balancer. Whenever a user opens a product page, the load balancer redirects the request to one of the size recommendation service systems. The size recommendation system fetches the user and product vectors from the database, computes the best fitting size and returns the same to the user. Our current system can handle 700K concurrent requests per minute. The online system is designed such that it can be horizontally scaled whenever traffic on the platform increases.

65% of users on our platform have multiple personas and buy for more than one user using the same account. To tackle this, we have built a profile level recommendation system where users can tag their current purchase to an existing profile or create a new profile. A user is allowed to have multiple profiles but a product can be tagged only to one profile. This gives us explicit feedback for whom the product was purchased. Each of these profiles are considered as different users by our model. The profile tagging system creates user vectors (profile vectors) in real time which are populated in the user vector database. Using the same size recommendation service we can now make size recommendations for new profile

Table 1: Precision scores for various articles

Article	Precision
Men Shirts	84.17%
Men Tshirts	84.33%
Women Dresses	82.07%
Women Kurtas	80.83%
Women Sweaters	93.8%
Men Trousers	95.25%

users. More than a million profiles were created in less than a week of the launch of profile tagging feature and essentially doubled the coverage of users. The profile vectors are fed back into the offline model creation pipeline to make the size recommendation system robust.

6 EXPERIMENTS AND RESULTS

In this section we provide data analysis and results of various experiments that were carried on our size recommendation system. We present both offline test results which were used to validate and test our method on historical data and online AB test results from our production systems.

6.1 Dataset

For the offline experiments we have used historical purchase data. As the goal of our model is to make future predictions, we have split the data by time and used 27 months' data as train data, 2 months' data as test data and 1 month's data for cross validation of model parameters. Our system builds a model for every article (eg : Men Shirts, Women Dresses, Men Trousers etc) on our platform. From here on we will consider the results of experiments carried on 'Men Jeans'. The total number of distinct styles for Men Jeans hosted on our platform is in the order of hundred thousands whereas the total transactions data in in millions. To protect our proprietary information, we do not report the actual product and transaction counts.

6.2 Data Analysis

To build our size recommendation models, we have collected product data and user transaction data to give us valuable insights about user behavior and buying trends. We have plotted the catalog size of product and transaction categories in Figures 5 and 6. Here, we have scaled the main dimension of products to US size convention of 4 to 22. We can see from Figure 5 that most of the product sizes of all four apparel categories lie within 10 to 16 sizes with women apparel categories (Women Sweaters, Women Sweatshirts) having more catalog inventory for smaller sizes in this range (more size 10 and size 12) than men apparel categories (Men Sweaters and Men Kurtas). This trend can also be seen in the transactions data in Figure 6. This can be attributed to the notion that most women are structurally smaller than men as the buying trends signify that popular sizes for women generally run a size or two smaller than men. This is reflected in the product inventory too as appropriate sizes have to be stocked for sale.

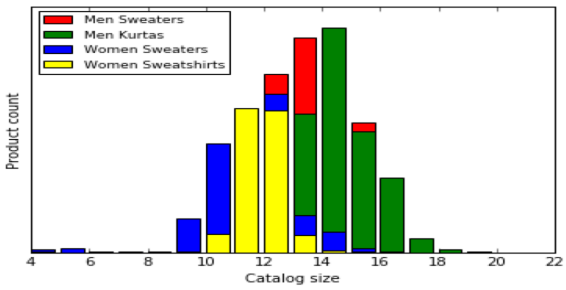


Figure 5: Product count vs catalog size

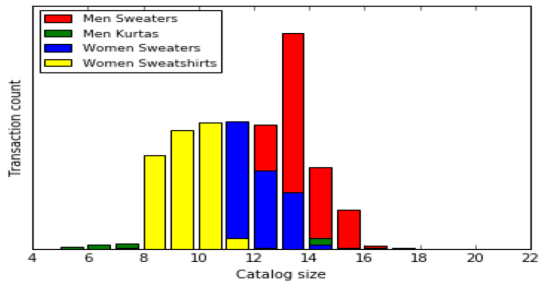


Figure 6: Transaction count vs catalog size

6.3 Model Results

We experimented with three different classifier models with various set of features. The features comprised of concatenated user and product vectors. For the first experiment we only used observable features in the classification task. Second experiment was performed using the latent features. For the third experiment we used both observable features and latent features for creating the corresponding product and user vectors. For comparison of results of the three experiments we have used Area under the curve in ROC. Figure 7 shows results for the three models. It can be seen from the figure that a model with a blend of observable and latent features is the most superior model. Table 1 shows the precision values for some articles and it is noticed that our approach generalizes well to different articles.

6.4 Local Cluster Results

In the previous sections we presented the problem of local clustering due to insufficiency of co-purchase data across various brands. Figure 8 compares different aggregation techniques; it can be seen that the median aggregation method performs better than mean aggregation method as the median metric is less susceptible to outliers. The concatenation aggregation method performs worse than the mean method as concatenated vectors are not robust. In the Doc2vec formulation, user vectors were created during training along with product vectors. This ensures that user vectors are a good representation of users in the size and fit space compared to any other aggregation method. The classification algorithm trained with Doc2Vec aggregation has the maximum AUC of 0.86.

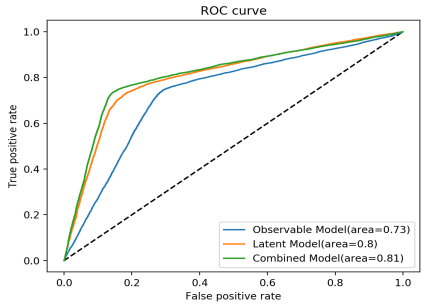


Figure 7: Comparison of classification models using ROC. It can be seen that the combined model with observable features and latent features is better than the individual models.

Table 2: Reduction in return rates for different articles

Article	Return rate reduction
Men Tshirts	6.4%
Men Jeans	8.4%
Men Trousers	4.8%
Men Sweatshirts	9.34%
Women Kurtas	1.08%

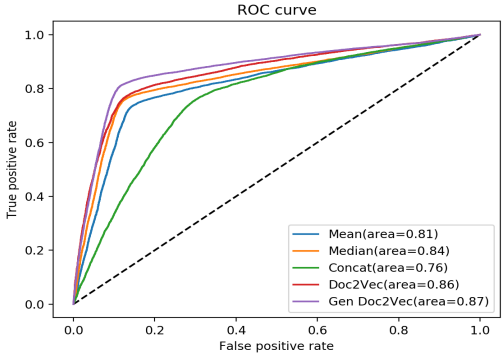


Figure 8: Comparison of various aggregation techniques for computing user vectors from product vectors. Doc2vec user vectors are better than all the other aggregation techniques. The purple curve shows Doc2vec user vector results with generated samples.

6.5 Online Results

Recommending correct sizes to users would result in reduction in returns due to size and fit issues. We set up an online A/B test to track the return reduction due to size and fit. In our platform we have a 30 day return policy which means users can return products purchased anytime in the next 30 days from the date of purchase. We conducted the test during a two-month window, wherein we considered the purchases made in the first month whereas the second month was a cool off period where we track the returns for these purchases. A/B test showed similar return rates for control and test sets before the start of the experiment. Members in the control set did not get any size recommendation while members from test set were shown recommendation from our system. Table 2 shows reduction in return rate for some articles for the month of December 2018.

Table 3: Test precision scores comparison for GBM approach and Matrix Factorization based approach

Article	GBM Test Precision	MF Test Precision
Men Shirts	84.17%	59.18%
Men Tshirts	84.33%	60.46%
Women Dresses	82.07%	60.73%
Women Kurtas	80.83%	58.53%
Women Sweaters	93.8%	61.36%
Men Trousers	95.25%	66.07 %

6.6 Implicit Matrix Factorization Results

We have used Implicit Matrix Factorization method outlined in Section 3.4 as the baseline method to compare our results. We ran both the methods on the test and train dataset for some of our article types and have tabulated the results in Table 3. Our approach outperforms the baseline approach by a huge margin for the test set. It was also observed that the baseline approach suffered from over fitting as the train set results were much higher than the test set results.

7 CONCLUSION

In this paper, we have described a size recommendation system which is used on our platform to guide users choose correct sizes while making an online purchase. We have introduced a method of creating latent features from purchase data and combining them with observable features to form a robust classification model for size recommendation. Further, we discussed the challenges which are unique to this domain like formation of local clusters in size and fit space and presented our approaches to overcome them. Due to inherent behaviour of multiple users purchasing from same account, we introduced a method of tagging orders to user-profiles which enabled us to make recommendation to multi-persona accounts with the same construct. To the best of our knowledge, this is the first attempt in the industry to scale size recommendation to multi-persona accounts. Although, we have demonstrated our work on fashion E-commerce data, it can be extended to other domains where we have sparse user and product signals. For our future work, we want to expand our work to make size recommendation available for users who have not made even a single purchase on our platform. We also intend to leverage the clothing contours of product images which might give valuable insights into product recommendations.

ACKNOWLEDGMENTS

The authors would like to thank Karthik M, Dheeraj Kumar, Sagar Grover, Srijit Sengupta and Gaurav Kumar for their contributions

in building data pipelines and providing valuable product insights in algorithm design, implementation and evaluation.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[2] Sagar Arora and Deepak Warrier. [n. d.]. Decoding Fashion Contexts Using Word Embeddings. ([n. d.]).

[3] Sandeep Bhanot and Srini R Srinivasan. 2013. A study of the Indian apparel market and the consumer purchase behaviour of apparel among management students in Mumbai and Navi Mumbai. *Navi Mumbai* (2013).

[4] AP Chan, J Fan, and WM Yu. 2005. Prediction of men's shirt pattern based on 3D body measurements. *International Journal of Clothing Science and Technology* 17, 2 (2005), 100–108.

[5] Chih-Hung Hsu. 2009. Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry. *Expert Systems with Applications* 36, 3 (2009), 4185–4191.

[6] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 263–272.

[7] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: a functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 129–138.

[8] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. 2000. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*. 512–518.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[11] Aaron Orendorff. 2019. The State of the Ecommerce Fashion Industry: Statistics, Trends & Strategy. Retrieved Jan 10, 2019 from https://www.shopify.com/enterprise/ecommerce-fashion-industry

[12] Vivek Sembium, Rajeev Rastogi, Atul Saroop, and Srujana Merugu. 2017. Recommending Product Sizes to Customers. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 243–250.

From Free-text User Reviews to Product Recommendation using Paragraph Vectors and Matrix Factorization

Georgios Alexandridis
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
gealexandri@islab.ntua.gr

Thanos Tagaris
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
thanos@islab.ntua.gr

Giorgos Siolas
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
gsiolas@islab.ntua.gr

Andreas Stafylopatis
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
andreas@cs.ntua.gr

ABSTRACT

Recent theoretical and practical advances have led to the emergence of review-based recommender systems, where user preference data is encoded in at least two dimensions; the traditional rating scores in a predefined discrete scale and the user-generated reviews in the form of free-text. The main contribution of this work is the presentation of a new technique of incorporating those reviews into collaborative filtering matrix factorization algorithms. The text of each review, of arbitrary length, is mapped to a continuous feature space of fixed length, using neural language models and more specifically, the Paragraph Vector model. Subsequently, the resulting feature vectors (the neural embeddings) are used in combination with the rating scores in a hybrid probabilistic matrix factorization algorithm, based on maximum a-posteriori estimation. The proposed methodology is then compared to three other similar approaches on six datasets in order to assess its performance. The obtained results demonstrate that the new formulation outperforms the other systems on a set of two metrics, thereby indicating the robustness of the idea.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Natural language processing**; **Maximum a posteriori modeling**; **Factorization methods**; **Learning latent representations**.

KEYWORDS

Recommender Systems; Collaborative Filtering; Matrix Factorization; Free-text reviews; Neural Language Processing; Word2Vec; Paragraph Vectors; Maximum A-posteriori Estimation

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316601

ACM Reference Format:
Georgios Alexandridis, Thanos Tagaris, Giorgos Siolas, and Andreas Stafylopatis. 2019. From Free-text User Reviews to Product Recommendation using Paragraph Vectors and Matrix Factorization. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316601

1 INTRODUCTION

The evolution of Collaborative Filtering (CF) Recommender Systems (RS) in recent years has resulted in a paradigm shift, moving away from systems that are solely based on the ratings' matrix to systems that incorporate user generated free-text reviews in the recommendation process as well. This reality has been spearheaded by the widespread adoption of recommendation modules in various social media platforms, which allow their users to evaluate items (venues, products, etc.) not only using traditional rating scales (e.g like/dislike, 1 to 5 star ratings) but also providing textual reviews along with other meta-information (Figure 1). This augmented preference data may be further analyzed through the application of text analysis and opinion mining techniques in order to extract various aspects of the conveyed information.

Review-based CF RS may be categorized, with respect to how the available reviews are populated, into *review-based user profile building* systems and *review-based product profile building* systems [7]. The former construct a user profile out of the available free-text reviews (their operational principle being similar to user-based CF approaches), while the latter create descriptive item profiles out of the same corpus of user-generated reviews (their operation being analogous to item-based CF methodologies). The proposed framework in this work, on the other hand, attempts to merge the aforementioned approaches by constructing profiles for *both* the items and the users.

Textual reviews may be processed using various methods, such as the creation of term-based profiles, employing standard techniques like the *term frequency-inverse document frequency* (TF-IDF) formula. Other methodologies populate free-text reviews in an effort to extrapolate preference scores in predefined, discrete scales, usually through the adoption of sentiment analysis tools. Of course, textual reviews may be incorporated in existing recommendation

algorithms as an additional data dimension that extends the ratings' matrix. In this setting, the user opinion or the topic extracted out of the respective reviews may be used as an extra weighting scheme, when computing user and item similarities.

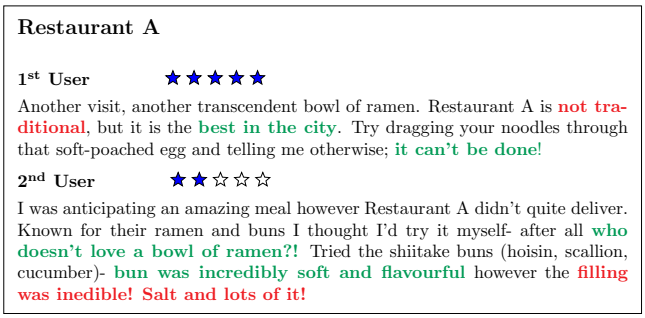


Figure 1: An example of free-text reviews of a restaurant

The majority of the aforementioned methodologies, however, place their emphasis on the word or the phrase level and many times fail to capture the whole context of the review. In Figure 1 for example, the phrase "not traditional" has a negative meaning, which is superseded by the next phrase ("best in the city"). The same review contains another phrase ("it can't be done!") that would have been labeled negatively by a sentiment analysis algorithm, but which, in fact, is used in a positive context. Weighting the appearances of positive and negative words and phrases is not very helpful either; e.g. the second review might contain much more positive than negative content but the last two phrases carry all the negative weight and justify the "poor" rating. Another interesting observation is that the phrase "Salt and lots of it!", which is used in an extremely negative context by the second reviewer, might fit a completely different narrative to another user who enjoys salty food.

Based on the reasoning above and extending earlier work [2], the approach outlined in this contribution is an efficient technique of enriching the user & item latent representations learned by CF-based ratings' matrix factorization algorithms, through the inclusion of the user-generated reviews into the procedure. This goal is achieved using a *neural language model* to map the free-text input of arbitrary size to a continuous real-valued numeric vector space. Unlike similar approaches that place the emphasis on the word or the phrase level, the novelty of the proposed system is that it learns the representations at the review level, using an appropriate model, the *Paragraph Vectors* (Section 3). In this way, it is possible to associate the same words and phrases, with a different meaning, to the various user and item profiles. This characteristic is exploited by *ParVecMF* (Section 5), a probabilistic matrix factorization algorithm applied to the combined information sources (ratings' matrix, vectorized representation of text) that learns the respective user & item latent vectors.

Additionally, Section 4 discusses matrix factorization methodologies, providing insight on the parameters and hyper-parameters involved and analyzing their influence on the produced user & item latent vectors. In Section 6, the strength of the proposed approach is assessed against three other techniques on a number of datasets and the obtained results (Section 7) on a set of two metrics and on three experiments ascertain the robustness of the overall framework. Finally, the paper concludes in Section 8, where possible future extensions are also discussed.

2 RELATED WORK

The incorporation of free-text user reviews into CF matrix factorization algorithms has been extensively studied in the relevant literature [7, 17]. However, most approaches attempt to extract the topic or the sentiment of textual reviews by examining the words they consist of, irrespective of the order they appear or their local context. In the most common scenario, each review is modeled as a distribution on a predefined number of latent ("hidden") topics and equivalently each topic is being defined by a distribution of words. This is the basis of the *Hidden Factors as Topics* (HFT) [14], the *Ratings Meet Reviews* (RMR), [12] and the *JMARS* [8] models.

Other approaches, like *TopicMF* [4], calculate the frequencies of words in user reviews and then construct a frequency matrix, which is factorized along with the ratings' matrix in a unified process. There are even cases where the *Paragraph Vector* model is also employed, like in the *Bag-of-Words regularized Latent Factor* model (BoWLF) [3]; yet the order of the words and their local context is not taken into account.

Some systems [1, 21] employ text parsers that perform a deep syntactic analysis of the reviews in order to discretize user sentiment with respect to various aspects of the items. These techniques do consider word order when extracting their context (e.g. positive, neutral, negative), but they do not encode the relation in-between words of the same context. Unlike neural language models, the aforementioned methodologies do not capture the writing *style* and therefore are not able to compute similarities and dissimilarities between free-text reviews in this dimension.

Neural language models, on the other hand, have been predominantly used in RS mainly as a means of achieving better representation of user preference in special cases (transactions [10], news stories [9], music tracks [22], check-ins at locations [18]).

Certain approaches, like *DeepCoNN++* [6] and *Collaborative Multi-Level Embedding* (CMLE) [25], do try to preserve both word order and word context. Nevertheless, they differ from *ParVecMF* in two key aspects; firstly, they employ pre-trained word embeddings on a general vocabulary (instead of training a specific NLP model for the given vocabulary) and secondly, they simply concatenate the embeddings of the words that make up the review, ignoring the *document* context. Finally, the *Language Model regularized Latent Factor* (LMLF) [3] model processes user reviews as a sequence of words, but it does not employ a neural language model for this task, using a recurrent neural network instead. To the best of our knowledge, our contribution is among the first to effectively account for word order & context, as well as document context at the same time, through the combination of paragraph vectors and CF matrix factorization in a unified learning approach.

3 THE PARAGRAPH VECTOR MODEL

The *Paragraph Vector* model [11] is an extension of the *Word2Vec* model [16] of distributed representations of words in a vector space.

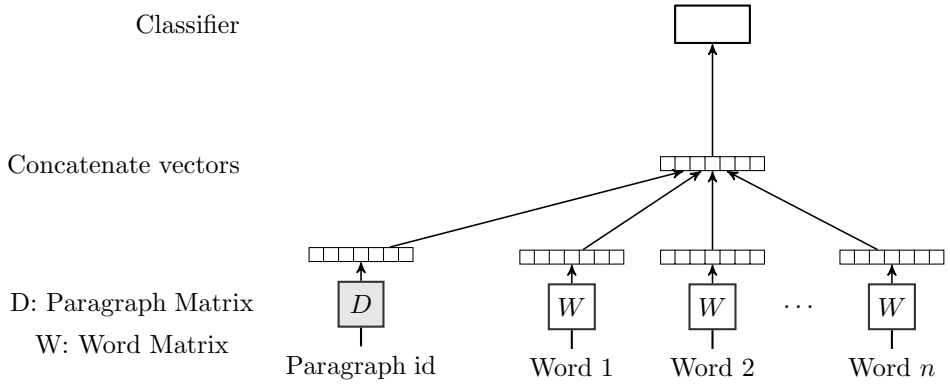


Figure 2: The Paragraph Vector model

In both models, every word is mapped to a unique vector, represented by a column in a word matrix W, indexed by the position of the word in the vocabulary.

Given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, the objective of both models is to maximize the average log probability of any given word in the sequence, conditioned on the appearance of the other words.

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_T|w_{t-k},\ldots,w_{t+k}) \tag{1}$$

Equation 1 above describes a prediction task, which is usually solved via the usage of a multiclass classifier such as *softmax*

$$p(w_T|w_{t-k},\ldots,w_{t+k}) = \frac{e^{y_{w_T}}}{\sum_i e_i^y}$$

Each term y_i is the un-normalized log-probability of each output word i, computed by

$$y = b + Uh(w_T|w_{t-k},\ldots,w_{t+k};W) \tag{2}$$

where U, b are the *softmax* parameters and function h is constructed by a concatenation (or averaging) of the word vectors extracted from W. For faster training, a *hierarchical softmax* with a *binary Huffman tree* structure is used in most cases [11]. Word vectors are usually trained via a neural network, using stochastic gradient descent and the gradient is obtained through backpropagation. After the training converges, the weights of each column of word matrix W correspond to the p dimensional vector representation of every vocabulary word.

The inspiration behind *Word2Vec* is that word vectors are asked to contribute to a prediction task about the next word in a sentence. In the *Paragraph Vector* model, paragraph vectors are used in a similar manner; they are asked to contribute to the prediction task of the next word given many contexts, sampled from the paragraph.

In addition to mapping each vocabulary word to a column in W, the *Paragraph Vector* model also maps every paragraph to a unique vector, represented by a column in a new document matrix D. The difference between *Paragraph Vectors* (Equation 2) and the *Word2Vec* model is that the output of the *softmax* classifier takes also into account the paragraph matrix D (Figure 2). Finally, the paragraph vector and word vectors are averaged (or concatenated) and are subsequently fed to the recommender system for the item prediction task.

A way to consider the paragraph token is as another word, which acts as a memory that remembers what is missing from the current context. This is equivalent to carrying the semantic information, or the topic, of the paragraph. For this reason, this model is called the *Distributed Memory* model of *Paragraph Vectors* (PV-DM). In this sense, the paragraph vectors, after the training phase, can be used simply as features for the paragraph.

Overall, the algorithm operates in two stages; the *unsupervised training* stage to get word vectors W and the *inference* stage to get paragraph vectors D. A third additional stage is to use D as a part of a probabilistic model for making predictions, like the one which is going to be presented in the subsequent section.

An important advantage of paragraph vectors is that they are learned from unlabeled data and thus can work well for tasks that do not have enough labeled data. They also address some of the key weaknesses of the bag-of-words models. Firstly, they inherit an important property of the *Word2Vec* model; the semantics of the words. In this space, we expect "orange" to be closer to "apple" than to "coffee". Secondly, they take into consideration the word order, at least in a small context, in the same way that an n-gram model with a large n would do.

Experimental results on two sentiment analysis tasks are presented in [11]. The first one is performed on the `Stanford Sentiment Treebank` dataset [19, 20], which has 11,855 sentences taken from the movie review site *Rotten Tomatoes*[1]. Every sentence in the dataset has a label which goes from very negative to very positive in the $[0-1]$ scale. The second one is performed on the `IMDB` dataset [13], consisting of 100,000 movie reviews with positive or negative labels. The authors report error rates that are by 16% and 15% better (relative improvement) than the best previous results on the two datasets, respectively.

The aforementioned advantages and the state-of-the-art results of the *Paragraph Vector* model have led us to choose this approach for the representation, in a vector space, of the textual information of our recommendation framework. However, it is used in a different context compared to the one described in [11]: rather than item descriptions, we analyze the items *reviews* made by the users. Then,

[1] https://www.rottentomatoes.com/

using this model, we try to extract the entire review's semantics with the ultimate goal of projecting similarly minded reviews in close distance in the vector space.

4 MATRIX FACTORIZATION

Matrix Factorization (MF) techniques approximate the mostly sparse and high dimensional ratings' matrix as the product of two other matrices, which are of lower dimensionality and more dense. More formally, given a positive semi-definite ratings' matrix $R \in \mathbb{R}_{+}^{n \times m}$ containing the ratings of n users on m items, the objective of the MF algorithm is to compute the elements of matrices $U \in \mathbb{R}^{n \times k}$ and $V^{\top} \in \mathbb{R}^{k \times m}$ so that their product $\widetilde{R}$ is as "similar" to R as possible, where similarity is defined in terms of a distance metric (e.g. the Euclidean distance).

$$R \approx \widetilde{R} \equiv UV^{\top} \tag{3}$$

The elements of matrix U (row vectors) may be regarded as the latent user factors that quantify the interests of each user. In a similar fashion, the elements of matrix $V^{\top}$ (column vectors) are regarded as the latent item factors that designate the opinion of the community about each item. It should also be noted that the length of the feature vectors (dimension k) is much smaller than both n and m ($k << n, m$).

In MF models, predictions are generated by multiplying the respective user and item latent vectors. That is, given a user i and an item j not rated by i, the predicted preference score $\widehat{r_{ij}}$ is equal to the inner product of the user latent vector $\mathbf{u}_i \in \mathbb{R}^{1 \times k}$ times the item latent vector $\mathbf{v}_j^{\top} \in \mathbb{R}^{k \times 1}$

$$\widehat{r_{ij}} = \mathbf{u}_i \mathbf{v}_j^{\top} \tag{4}$$

The most popular MF technique used in RS is the *Alternating-Least-Squares with Weighted-λ-Regularization* (ALS) algorithm [26]. In this approach, learning the latent representations of the user & item features is expressed as a minimization problem

$$\min_{\mathbf{u}_i, \mathbf{v}_j^{\top}} \left[\sum_{(i,j) \in R} (r_{ij} - \mathbf{u}_i \mathbf{v}_j^{\top})^2 + \lambda \left(||\mathbf{u}_i||^2 + ||\mathbf{v}_j||^2 \right) \right] \tag{5}$$

for all user and item pairs (i, j) in R (please note that a slightly different notation than in [26] is used).

The optimization problem described in Equation 5 above is non-convex since both $\mathbf{u}_i, \mathbf{v}_j^{\top}$ are unknown. However, if one of the unknowns is fixed to a constant value, then the same problem becomes quadratic and can be solved by a least-squares method. The ALS algorithm, at each iteration, interchangeably fixes one variable while optimizing the other, until convergence is achieved. In order to avoid overfitting, the magnitudes of the obtained solutions for $\mathbf{u}_i, \mathbf{v}_j^{\top}$ are penalized (regularized) by a certain λ term, which constitutes a hyper-parameter of the approach.

4.1 Probabilistic Matrix Factorization

Probabilistic Matrix Factorization (PMF) is a special case of MF in which the underlying assumption is that the elements of matrices U, V stem from a probability distribution of a known type but of unknown parameters. The most common choices are the Normal and the Poisson distributions. In case of the former, the element r_{ij} of the ratings' matrix is thought to originate from the Normal distribution with a mean value of $\mathbf{u}_i \mathbf{v}_j^{\top}$ and a constant specificity of c (commonly set to $c = 1$ [26])

$$r_{i,j} \sim \mathcal{N}(\mathbf{u}_i \mathbf{v}_j^{\top}, c^{-1}) \tag{6}$$

Vectors $\mathbf{u}_i, \mathbf{v}_j$ constitute the parameters of a system described by the data r_{ij} and therefore can be approximated via the standard *Bayesian inference rule*

$$P(\mathbf{u}_i, \mathbf{v}_j | r_{ij}) \propto P(r_{ij} | \mathbf{u}_i, \mathbf{v}_j) \times P(\mathbf{u}_i) \times P(\mathbf{v}_j) \tag{7}$$

where $P(\mathbf{u}_i), P(\mathbf{v}_j)$ designate the *a-priori* distributions of the parameters. The system described in Equation 7 may be learned using *Maximum A-posteriori* (MAP) estimation, which tries to maximize the logarithm of the *posterior*

$$\begin{aligned} \arg\max_{\mathbf{u}_i, \mathbf{v}_j} \log P(\mathbf{u}_i, \mathbf{v}_j | r_{ij}) \quad &\propto \quad \arg\max_{\mathbf{u}_i, \mathbf{v}_j} \log P(r_{ij} | \mathbf{u}_i, \mathbf{v}_j) \\ &+ \quad \arg\max_{\mathbf{u}_i} \log P(\mathbf{u}_i) \\ &+ \quad \arg\max_{\mathbf{v}_j} \log P(\mathbf{v}_j) \end{aligned} \tag{8}$$

5 PARVECMF

In the proposed architecture, the main objective is to combine the *Paragraph Vector* model (Equation 1), which maximizes the log-likelihood of the appearance of any word in the text, conditioned on its surrounding words, and maximum a-posteriori estimation (Equation 8), which maximizes the log-likelihood of a specific user evaluating a particular item with the specified rating score, given the respective user & item latent vectors. Since the two objectives are different, a methodology should be devised in order to integrate them in a unified optimization scheme.

Initially, the free-text reviews are processed by the *Paragraph Vector* model, in two independent learning procedures; one that accumulates the reviews on a per user basis and another that accumulates them on a per item basis. In either case, the context of an individual textual review is determined by the respective rating score. The advantages of this approach are twofold; on the item level, the system amasses the opinion of the community for the item at-hand, along with higher-level characterizations that may have not been included in the item's description in the first place. On the user level, the system learns the "vocabulary" of each user (the words/phrases used when positive or negative feelings are expressed) along with each user's particular taste (recall the discussion on the salty food in Section 1).

Subsequently, and inspired by similar joint optimization approaches (e.g. [23], [25]), the elements of the user latent vectors are assumed to originate from the neural embeddings of the user's reviews ($\boldsymbol{\theta}_i \in \mathbb{R}^k$) plus and offset ($\boldsymbol{\epsilon}_i \in \mathbb{R}^k$) that aggregates the two different optimization objectives. In practice, this offset models the relationship between the user's rating and reviewing behavior

$$\begin{aligned} \mathbf{u}_i &= \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, \lambda_i^{-1} I_k) \end{aligned} \tag{9}$$

The offset ($\boldsymbol{\epsilon}_i$) is also assumed to originate from the multivariate normal distribution with a mean vector of $\mathbf{0} \in \mathbb{R}^k$ and a diagonal covariance matrix of $\lambda_i^{-1} I_k \in \mathbb{R}^{k \times k}$. The covariance matrix is chosen to be diagonal because, by definition, in MF the features are independent of one another.

In a similar manner, the latent vector for item j is derived from the neural embeddings of all the reviews it has received ($\boldsymbol{\theta}_j \in \mathbb{R}^k$), plus an offset ($\boldsymbol{\epsilon}_j \in \mathbb{R}^k$) that models the relationship between the ratings of the item and the said reviews

$$\begin{aligned} \mathbf{v}_j &= \boldsymbol{\theta}_j + \boldsymbol{\epsilon}_j \\ \boldsymbol{\epsilon}_j &\sim \mathcal{N}(\mathbf{0}, \lambda_j^{-1} I_k) \end{aligned} \tag{10}$$

Since the neural embeddings $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j$ described in Equations 9-10 above are learned by the *Paragraph Vector* model in a separate process, they are considered to be part of the data of the model (along with the rating value r_{ij}). Therefore, the Bayesian inference rule of Equation 7 becomes

$$P(\mathbf{u}_i, \mathbf{v}_j | r_{ij}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \propto P(r_{ij} | \mathbf{u}_i, \mathbf{v}_j) \times P(\mathbf{u}_i | \boldsymbol{\theta}_i) \times P(\mathbf{v}_j | \boldsymbol{\theta}_j) \tag{11}$$

Combining Equations 6 through 11 and re-writing the offsets $\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j$ as the difference between the user & item latent vectors $\mathbf{u}_i, \mathbf{v}_j$ and their respective neural embeddings $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j$, yields

$$P(\mathbf{u}_i, \mathbf{v}_j | r_{ij}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \propto \mathcal{N}(\mathbf{u}_i \mathbf{v}_j^\top, c) \times \mathcal{N}(\mathbf{0}, \lambda_i^{-1} I_k) \times \mathcal{N}(\mathbf{0}, \lambda_j^{-1} I_k) \tag{12}$$

and the log-likelihood $\mathcal{L}$ of the a-posteriori probability becomes

$$\begin{aligned} \mathcal{L} = &- \sum_{i=1}^{n} \frac{\lambda_i}{2} (\mathbf{u}_i - \boldsymbol{\theta}_i) I_k (\mathbf{u}_i - \boldsymbol{\theta}_i)^\top \\ &- \sum_{j=1}^{m} \frac{\lambda_j}{2} (\mathbf{v}_j - \boldsymbol{\theta}_j) I_k (\mathbf{v}_j - \boldsymbol{\theta}_j)^\top \\ &- \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{c}{2} (r_{ij} - \mathbf{u}_i \mathbf{v}_j^\top)^2 \end{aligned} \tag{13}$$

The MAP estimates for the optimal values of the elements of the user & item latent vectors are located at those points where the gradient of $\mathcal{L}$ with respect to the parameters of the model is equal to zero, leading to the following update rules

$$\mathbf{u}_i \leftarrow \left[c \mathbf{r}_i V + \lambda_{\mathbf{u}_i} \boldsymbol{\theta}_i\right] \left[c V^\top V + \lambda_{\mathbf{u}_i} I_k\right]^{-1} \tag{14}$$

$$\mathbf{v}_j \leftarrow \left[c \mathbf{r}_j U + \lambda_{\mathbf{v}_j} \boldsymbol{\theta}_j\right] \left[c U^\top U + \lambda_{\mathbf{v}_j} I_k\right]^{-1} \tag{15}$$

The graphical model of the system is illustrated in Figure 3.

An interesting observation on Equations 14-15 above is that the magnitude of regularization hyper-parameters λ_u and λ_v determine the extend to which the free-text reviews (in the form of the computed neural embeddings) contribute to the formation of the user & item latent vectors. Setting any (or both) of them to zero would result in the respective user & item latent vectors to reflect the information contained in the ratings' matrix only. On the other hand, if λ_u, λ_v assume very large values, then the effect of the ratings in the factorization process will be diminished and the learned latent vectors will be solely based on the representation of the free-text user reviews in the document vector space. Obviously, when $\lambda_u = \lambda_v = 1$ the ratings and the textual reviews have an equal impact on the optimization process.

6 EXPERIMENTS

6.1 Datasets

A set of experiments has been performed on number of datasets selected from two distinct collections. The first one is the `Amazon product data` collection of datasets [15], extracted from Amazon

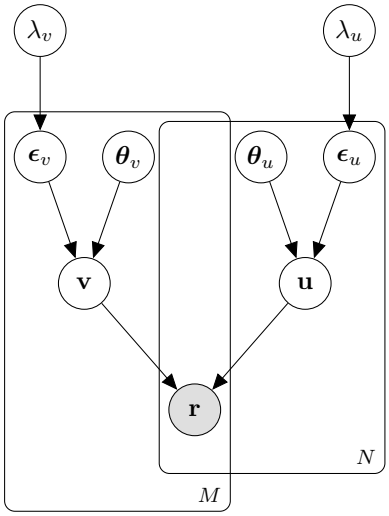


Figure 3: The graphical model of the proposed system

over a period spanning almost 20 years, while the second one is the `Yelp Open Dataset` [24], which contains more than 5,200,000 reviews on about 174,000 businesses from 11 metropolitan areas along with more than 200,000 pictures, provided by the respective online service.

Table 1: Dataset characteristics

Dataset	Users	Items	Reviews	M.w.+
A. `Amazon product data`				
1. `Movies and TV`	123,960	50,052	1,697,533	86
2. `Electronics`	192,403	63,001	1,689,188	64
3. `CDs and Vinyl`	75,258	64,443	1,097,592	119
4. `Kindle Store`	68,223	61,934	982,619	62
5. `Home and Kitchen`	66,519	28,237	551,682	60
B. `Yelp Open Dataset`				
1. `Yelp Reviews`*	213,170	94,303	3,277,932	88

+ Median words per review, * reduced to 5-core

Table 1 outlines the characteristics of the datasets used in the experiments, while Table 2 summarizes the available fields of each review entry in every dataset. Five of the largest datasets in the `Amazon product data` collection were chosen, along with the reviews subset of the `Yelp Open Dataset`. It should be noted that the Amazon datasets are in the form of 5-core; that is, ever user/item in the datasets has contributed/received at least 5 reviews. In order for the results to be comparable, the Yelp dataset has also been reduced to 5-core.

As it is evident from the aforementioned tables, each review is very descriptive in terms of the available information and meta-information. Nevertheless, the datasets remain extremely sparse, having all of them a data density of $0,01\% - 0,03\%$ (computed using the standard ratio of the number of reviews over the product of the number of users times the number of items) and the distribution

Table 2: Available fields per review

Field		Description
Amazon	Yelp	
asin	business_id	The unique item id
reviewerID	user_id	The unique user id
reviewerName	name	The user's name
helpful	useful	The ratio of users who found this review useful (Amazon) & the Number of useful votes (Yelp)
overall	stars	The rating of the product (5-star scale)
reviewTime	date	The time of the review
summary	-	Short summary of the review, written by the reviewer himself
reviewText	text	The actual text of the review
-	funny	Number of funny votes this review received
-	cool	Number of cool votes this review received

of the number of reviews over the number of users and items has *long-tail* characteristics [5].

6.2 Procedure

The sparsity of the data and the other properties of the datasets are reflected on certain experimental design choices. As the both of them are eventually in the form of 5-core, the recommendation list size is fixed to 5 items. The relatively small number of median words per review (5^{th} column of Table 1) also means that all words are to be considered by the *Paragraph Vector* model, therefore the minimum word frequency is set to one. The remaining parameters of the *Paragraph Vector* model (window size and number of epochs) are set to some widely-adopted predefined values (5 and 10 respectively).

In order to assess the effectiveness of *ParVecMF*[2], it is compared to three other systems; *ALS* [26] (Section 4), a pure matrix factorization approach, *HFT*[3] [14] (Section 2), which incorporates the free-text reviews in MF, modeling each word irrespective of the others as a distribution of topics and finally *DeepCoNN++*[4] [25] (Section 2), which preserves word order and context but does not take into account the document (review) context. The hyper-parameters of the other three approaches have been set to the optimal values reported by their respective authors, with the exception of the pre-trained vectors used in *DeepCoNN++*, which were chosen to be of the same dimensionality as the paragraph vectors in order for the comparisons to be fair.

6.3 Metrics

The performance of all systems is measured on two rank metrics, an information-retrieval one (*Mean Average Precision at N* or MAP@N) and an accuracy one (*Mean Reciprocal Rank at N* or MRR@N). Both metrics are defined over the total number of recommended lists of size N for the users in the test set T. Mean average precision is the mean of average precision for each recommended list

$$MAP = \frac{1}{|T|}\sum_{u=1}^{|T|}\overline{Pr(u)}$$

where the average precision in a list of length L for a user u is defined as

$$\overline{Pr(u)} = \frac{1}{|I_u|}\sum_{i=1}^{L} Pr(i) \times rel(i)$$

with I_u being the set of relevant items for user u, $Pr(i)$ the precision at cut-off i in the list and $rel(i)$ is equal to one if the i-th item in the list is relevant for u (and zero otherwise).

The mean reciprocal rank is defined as

$$MRR = \frac{1}{|T|}\sum_{l=1}^{|T|}\frac{1}{\text{rank}_l}$$

where rank_l designates the position of the first relevant item for the l-th user in the recommendation list. For both metrics and for the datasets at hand, the relevance threshold is set to four out of five stars. It should also be noted that MAP and MRR are computed for the "good" items in the test set (those items in the test set above the relevance threshold). Finally, the evaluation methodology followed is 5-fold cross validation computed over the available evaluations.

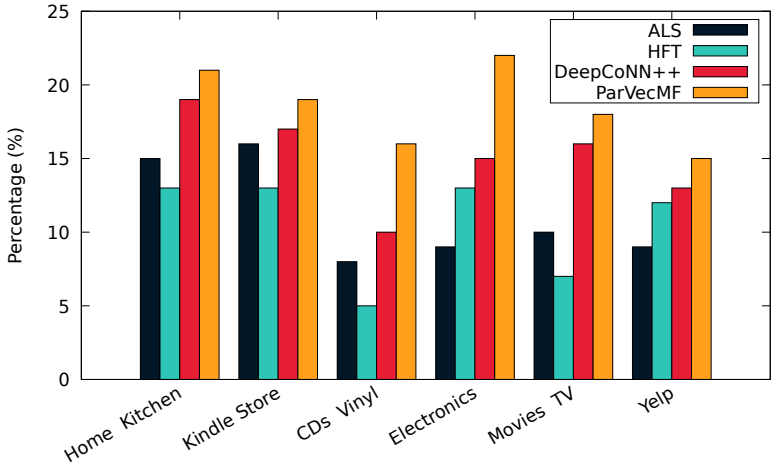
7 RESULTS

Figures 4-6 summarize all experimental results. All metrics are averaged over the 5 folds, with the statistical significance of all results being asserted by the *Wilcoxon singed rank test* ($p_{value} < 0.01$), for all possible model pairs on all datasets. More specifically, Figure 4 displays the results of the evaluation procedure for latent user/item vectors of 50 features and for a recommendation list of 5 items.

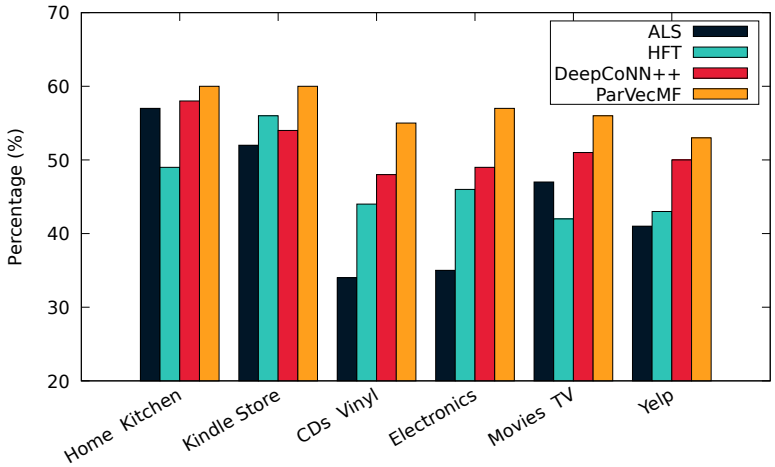
Figure 4a exhibits the results of the MAP metric. As it is evident, *ParVecMF* outperforms the other approaches on all datasets, proving the robustness of the methodology. An interesting observation is that the pure MF algorithm (*ALS*) performs better than the simpler approach (*HFT*) on most datasets, an indication that words should not be processed irrespective of their local context. Additionally, *DeepCoNN++* is surpassed by our system on all datasets. This behavior is attributed to two factors; firstly, *DeepCoNN++* takes into consideration only the word context and it misses the review context, as it merges all reviews of each user/item in its input, regardless of their actual polarity. Secondly, it is based on pre-trained vectors of a general vocabulary instead of training an NLP model on the specific vocabulary used in each dataset.

Figure 4b summarizes the performance of all examined systems on the MRR metric. Again, the proposed system exhibits a clear performance lead, which means that it is, on average, more capable of recommending at least one interesting item to the users. With the exception of the `Home & Kitchen` dataset, this conclusion also holds true for the other two systems that process free-text reviews (*HFT*, *DeepCoNN++*) when compared to the pure MF approach (*ALS*).

[2]https://git.islab.ntua.gr/gealexandri/parvecmf
[3]http://cseweb.ucsd.edu/ jmcauley/code/code_RecSys13.tar.gz
[4]https://github.com/chenchongthu/DeepCoNN



(a) Mean Average Precision for a list of 5 items (MAP@5)



(b) Mean Reciprocal Rank for a list of 5 items (MRR@5)

Figure 4: Results on all datasets (50 features)

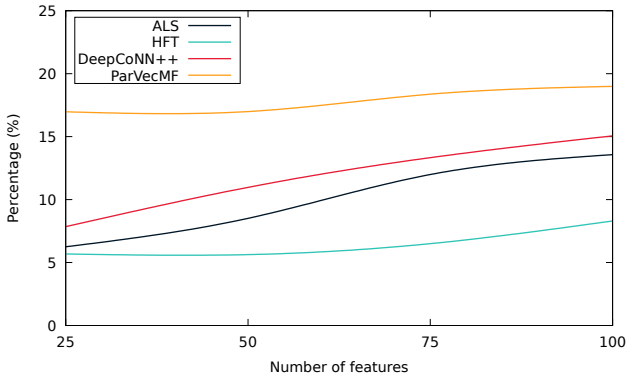
Therefore, including textual reviews as an extra input dimension generally boosts RS performance.

In another experiment, a dataset of medium size has been selected (`Cds and Vinyl`) and the algorithms are evaluated on the chosen set of metrics for a varying length of the latent user & item vector space. Figure 5a illustrates the increase of the MAP metric for all systems examined, as the length of the latent user & item vectors grows. The performance of *ALS* and *DeepCoNN++* is more dependant on the size of the feature space, while the proposed approach is robust even at smaller feature spaces. *HFT* efficiency, on the other hand, remains low at small feature spaces and starts to gradually increase only for larger spaces. In Figure 5b, the respective change of the MRR metric with respect to the feature space size is demonstrated. It is obvious that the ability of all algorithms to produce at least one meaningful recommendation is greatly affected by the length of the feature space and the larger it becomes, the more informative the obtained user/item latent features are.

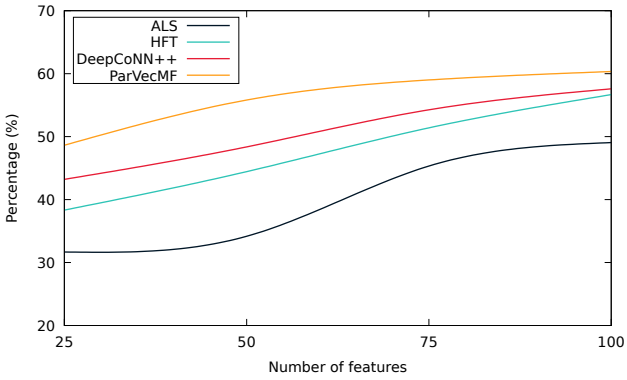
Finally, in Figure 6, the effect of the regularization parameters λ_u and λ_v, which determine the extend to which the free-text reviews contribute to the user & item latent vectors, is assessed on *ParVecMF* for the `Home and Kitchen` dataset. Each time, one of λ_u, λ_v is fixed to 1 and the value of the other is alternated between very small (0.01) and large (10) values. When λ_u, λ_v are individually set to a very small value, it means that the respective user/item neural embeddings play a minimal role in the optimization process of Section 5 and when they are set to a large value, then they dominate the said process. As expected, the performance of both metrics deteriorates in the latter case, meaning that no specific component of the proposed framework should overshadow the other.

8 CONCLUSION

This work outlines a new methodology of combining user reviews, in the form of neural embeddings, and ratings in probabilistic matrix

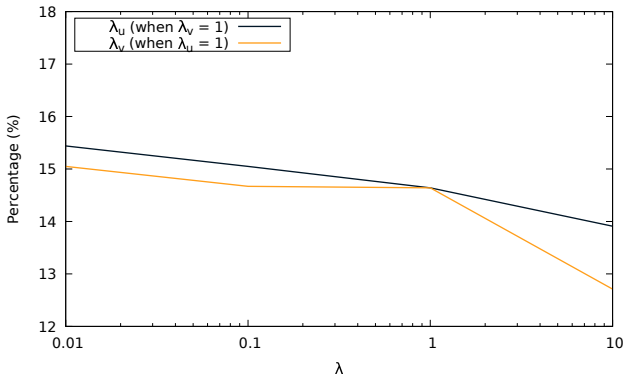


(a) Mean Average Precision for a list of 5 items (MAP@5)

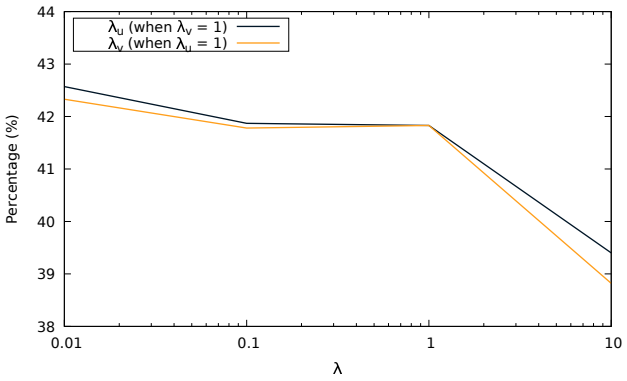


(b) Mean Reciprocal Rank for a list of 5 items (MRR@5)

Figure 5: The effect of the feature space on all systems (`CDs and Vinyl` dataset)



(a) Mean Average Precision for a list of 5 items (MAP@5)



(b) Mean Reciprocal Rank for a list of 5 items (MRR@5)

Figure 6: The effect of the regularization parameters on *ParVecMF* (`Home and Kitchen` dataset)

factorization. The presented results on a variety of datasets support the robustness of the proposed method and clearly indicate that this is a research direction worth of further exploring. The main novelty of our contribution is the use of the *Paragraph Vector* model as a means of representing textual reviews given by the users themselves instead of encoding item descriptions or general user preference data, like other approaches do.

The usefulness of this design choice for recommender systems should be once again stressed, as similarities in the way items are reviewed by the users imply an affinity in taste. The *Paragraph Vector* model further stimulates this notion of likeness, as it expands the similarity in the context of reviews that consist of different words. This domain is ignored by representations based solely on the word level, either dependent on (*DeepCoNN++*) or independent of (*HFT*) word order. Additionally, the importance of training an NLP model on each specific domain in order to extract the respective semantics, in contrast to using pre-trained vectors on general vocabularies (*DeepCoNN++*), should also be highlighted.

In general, review-based recommender systems are expected to become more and more popular, as the amount and dimensions of the available information evolves with time. This fact means that neural language models are likely to find even more applications in recommender systems, especially in conjunction with domains such as the social networks.

Finally, the effectiveness of the presented model should be further assessed by performing more comparative experiments to other related approaches. It would also be interesting to extend the proposed framework in order to incorporate more information and meta-information (e.g.the helpfulness of each review), where available (Table 2).

REFERENCES

[1] S. Aciar, D. Zhang, S. Simoff, and J. Debenham. 2007. Informed Recommender: Basing Recommendations on Consumer Product Reviews. *IEEE Intelligent Systems* 22, 3 (May 2007), 39–47. https://doi.org/10.1109/MIS.2007.55

[2] Georgios Alexandridis, Georgios Siolas, and Andreas Stafylopatis. 2017. ParVecMF: A Paragraph Vector-based Matrix Factorization Recommender System. *CoRR* abs/1706.07513 (2017). arXiv:1706.07513 http://arxiv.org/abs/1706.07513

[3] Amjad Almahairi, Kyle Kastner, Kyunghyun Cho, and Aaron Courville. 2015. Learning Distributed Representations from Reviews for Collaborative Filtering. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. ACM, New York, NY, USA, 147–154. https://doi.org/10.1145/2792838.2800192

[4] Yang Bao, Hui Fang, and Jie Zhang. 2014. TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*. AAAI Press, 2–8.

[5] Òscar Celma. 2010. *The Long Tail in Recommender Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 87–107. https://doi.org/10.1007/978-3-642-13287-2_4

[6] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. 1583–1592.

[7] Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction* 25, 2 (2015), 99–154. https://doi.org/10.1007/s11257-015-9155-5

[8] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 193–202. https://doi.org/10.1145/2623330.2623758

[9] Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic, and Narayan Bhamidipati. 2015. Hierarchical Neural Language Models for Joint Representation of Streaming Documents and Their Content. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 248–255. https://doi.org/10.1145/2736277.2741643

[10] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in Your Inbox: Product Recommendations at Scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1809–1818. https://doi.org/10.1145/2783258.2788627

[11] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Eric P. Xing and Tony Jebara (Eds.), Vol. 32. PMLR, Bejing, China, 1188–1196. http://proceedings.mlr.press/v32/le14.html

[12] Guang Ling, Michael R. Lyu, and Irwin King. 2014. Ratings Meet Reviews, a Combined Approach to Recommend. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 105–112. https://doi.org/10.1145/2645710.2645728

[13] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 142–150. http://dl.acm.org/citation.cfm?id=2002472.2002491

[14] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 165–172. https://doi.org/10.1145/2507157.2507163

[15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2766462.2767755

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

[17] Michael P. O'Mahony and Barry Smyth. 2018. *From Opinions to Recommendations*. Springer International Publishing, Cham, 480–509. https://doi.org/10.1007/978-3-319-90092-6_13

[18] Makbule Gulcin Ozsoy. 2016. From word embeddings to item recommendation. *arXiv preprint arXiv:1601.01356* (2016).

[19] Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 115–124. https://doi.org/10.3115/1219840.1219855

[20] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Vol. 1631. 1642.

[21] Anna Stavrianou and Caroline Brun. 2015. Expert Recommendations Based on Opinion Mining of User-Generated Product Reviews. *Computational Intelligence* 31, 1 (2015), 165–183. https://doi.org/10.1111/coin.12021

[22] Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2016. Meta-Prod2Vec: Product Embeddings Using Side-Information for Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 225–232. https://doi.org/10.1145/2959100.2959160

[23] Chong Wang and David M. Blei. 2011. Collaborative Topic Modeling for Recommending Scientific Articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 448–456. https://doi.org/10.1145/2020408.2020480

[24] Yelp. 2018. The Yelp Open Dataset. https://www.yelp.com/dataset/

[25] Wei Zhang, Quan Yuan, Jiawei Han, and Jianyong Wang. 2016. Collaborative Multi-level Embedding Learning from Reviews for Rating Prediction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2986–2992. http://dl.acm.org/citation.cfm?id=3060832.3061039

[26] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. 2008. Large-Scale Parallel Collaborative Filtering for the Netflix Prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management (AAIM '08)*. Springer-Verlag, Berlin, Heidelberg, 337–348. https://doi.org/10.1007/978-3-540-68880-8_32

Exploration of Product Search Intents via Clustering of Query Clusters

Omar Alonso
Microsoft
omalonso@microsoft.com

Vasileios Kandylas
Microsoft
vakandyl@microsoft.com

Serge-Eric Tremblay
Microsoft
sergetr@microsoft.com

ABSTRACT

We describe a system that organizes search results in the context of an exploratory product search session where the user is researching goods. Compared to existing approaches that use predefined categories to filter results by attributes, we organize information needs based on queries instead of documents. The idea is to organize queries around the same topic and produce a hierarchical representation of intents that describe information about a product from different perspectives. We present a prototype implementation using a real-world data set of 24M queries.

ACM Reference Format:
Omar Alonso, Vasileios Kandylas, and Serge-Eric Tremblay. 2019. Exploration of Product Search Intents via Clustering of Query Clusters. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316606

1 INTRODUCTION

Product search continues to be a very popular activity as most products can be purchased online. Users can use a web search engine or a shopping website to locate specific merchandise. Web search engines return product information as a list of pages ranked by relevance, whereas a shopping site organizes relevant results according to categories and past purchase behavior. We propose an alternative based on the queries that users enter when they are looking for a product and how that data can be used for organizing the various aspects that are required for making a purchase decision. We hypothesize that searching for a product involves more than sifting through pages, in particular when items are somewhat pricey, or the user is not familiar with the product line.

Consider the following scenario: we are looking to purchase a new dishwasher to replace the old one. We don't know a lot about this type of appliances and we need to do a bit of research. What is the latest on this product? What is a good brand? Are there any reviews? A search on a shopping vertical shows a wide range of products with many filters that are useful when we are ready to make a purchase decision, which is not our case yet. Web search results do cover more content but, at the same time, not fully related content is included. When people are looking for a dishwasher, what are the most common things that they express?

We are interested in organizing product information for supporting exploratory search scenarios [6]. Our work centers around structure and organization and not on user interfaces. In fact, we use a standard clustering search interface to show search results.

There is previous research on systems that organize search results and how to present them in a user interface so the user can interact with the collection in different ways. The book by Hearst describes interfaces for grouping search results and collections by flat categories, hierarchical categories, and faceted categories [4].

A well-known approach for organizing content is clustering, a technique that derives groups and labels in an unsupervised fashion. A disadvantage of clustering is that the output quality is not as good as a category system. To avoid some of the problems with traditional clustering on documents, web clustering organizes items that are returned by a web search engine using as input the titles and snippets of the returned pages [3]. There is recent work on bundling search results using composite retrieval for exploratory entity search task using a question-answering data set [2].

While product search is a pillar in e-commerce, published research on this topic is sparse. A high-level overview of the techniques used by Amazon is presented in [9] where the authors identify blending of results from different categories as a big challenge for ranking. Li et al. [5] points out that the process of buying a product is different than the process of locating relevant documents in a traditional search engine and propose a system that recommends products that have the best value.

Web search engines do have a related searches feature, a short list of queries, that is usually placed at the end of the page. We are interested in expanding the notion of related queries in combination with clustering to organize topic exploration for product search. In contrast to previous work that clusters web search results or a set of documents, we pre-process a search engine query log by *clustering queries* using clicks. In the same spirit as web clustering, we search the query clusters and then cluster the search results of such queries. That is, an offline clustering step and then an on-the-fly clustering, or bundling, to organize the output of the original clusters according to the user queries. Our goal is to provide a summary of the most relevant intents that users provide in the context of a product search session. Figure 1 shows the system high level architecture.

2 METHODS

We now describe the two main techniques implemented: clustering queries based on clicks and packaging the clusters using a suffix array-based algorithm.

2.1 Clustering Queries

In order to infer semantic connections between queries from the search query log, we exploit links clicked for each keyword. This approach lets us detect non-obvious semantic associations, and it

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316606

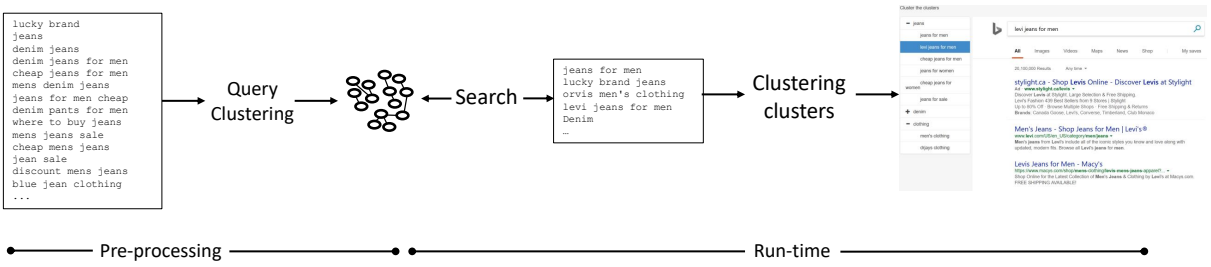


Figure 1: System architecture. The data processing phase consists of clustering a search query log using clicks and producing a data sets of clusters. At run-time, the search results for a query are clustered and presented as a hierarchical tree for the user. The user can explore the many aspects by navigating the tree that is presented on the left.

is practical to implement [1]. Consider a vector space where each dimension represents a link from the Bing search query logs. We associate each query to a vector where each component of the vector represents the number of clicks on the link. These vectors are high-dimensional but very sparse. We compute cosine similarity between the two vectors to obtain similarity. If we compute the distance between every possible pair of queries, we obtain a query similarity graph. In this weighted, undirected graph, each vertex represents a query, and the edges describe their similarity.

Once the query similarity graph is built, the next step is to create groups of related queries with community detection. The idea is to identify communities (clusters) of queries which are densely connected to each other, but loosely connected to the rest of the graph using modularity maximization [7].

Consider an undirected graph $G = (V, E)$. For the sake of presentation, we consider that this graph is not weighted, but that more than one edge can connect two nodes. Consider a set of vertices $C \in V$. The modularity measures how densely connected C is. To compute it, we count the number of edges within the set, and compare to what we would expect if the edges were drawn randomly between G's vertices, preserving the vertex degrees. The modularity is the difference between these two terms. Let m_C be the number of edges and $E[m_C]$ the expected number of edges in the set C. The modularity of C is:

$$Mod(C) = m_C - E[m_C]$$

If the vertices of the graph G are partitioned into p partitions $C_1, ..., C_p$, then the total modularity $TMod$ is the sum of the modularities of each of these partitions. To find the modularity of a set of vertices C, we can compute m_C by counting the number of edges within C. But how do we obtain the expected number $E[m_C]$? Let's draw an edge at random between two vertices of G. Let P_C describe the probability that the edge connects two vertices of C, and let m_G denote the number of edges in the graph. We obtain:

$$E[m_C] = m_G * P_C$$

We now compute the probability P_C. Let $D_G = 2 * m_G$ represent the sum of all the degrees of all the vertices of the graph, and let D_C represent the sum of all the degrees of the vertices in C. For a given edge, the probability that one of the endpoints ends up in the set C is D_C/D_G. Therefore,the probability of having both endpoints in the community is:

$$P_C = (D_C/D_G)^2$$

Putting everything together, we obtain:

$$Mod(C) = m_C - m_G * (D_C/D_G)^2$$

For clustering via modularity maximization we proceed as follows. We initialize the algorithm by assigning each vertex to its own community. Then, at each iteration, we find the two closest communities, and merge them. We stop when we cannot improve the score anymore, or when we have reached a satisfying number of communities. Given the size of the query logs, we use a custom variant of Newman's procedure presented in [8] that includes the following three steps:

(1) For each community, list all the neighbor communities. Two communities are neighbors if (a) they are connected and (b) if we union them, the total modularity increases. We obtain several neighborhoods, one for each community.
(2) The neighborhoods found in Step 1 are overlapping: one community may belong to several neighborhoods. To remediate this, take each community, list all the neighborhoods to which it belongs and keep the closest one (modularity increase is as large as possible).
(3) For each neighborhood, aggregate all the communities into one large, new community.

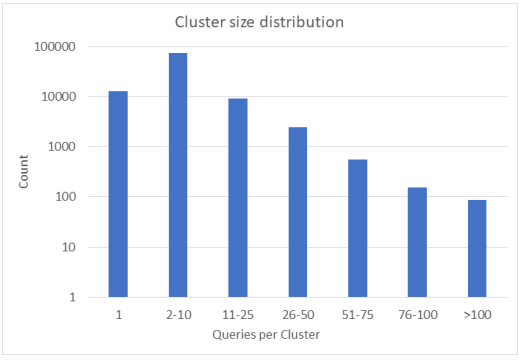


Figure 2: Cluster size distribution.

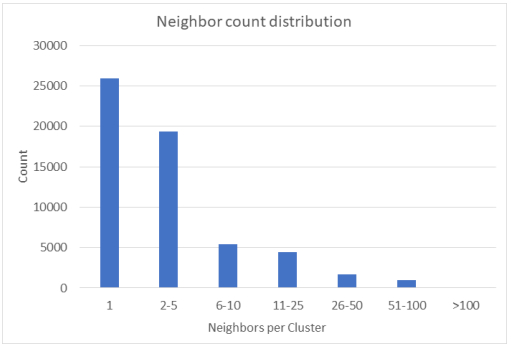


Figure 3: Cluster neighbors.

Figure 2 shows community (cluster) sizes. We observe that a large majority of communities contains between 2 and 10 queries, and that there are very few communities with more than 50 items. After the communities have been detected, we also find the neighboring communities, i.e. those communities which are close (up to a threshold) but not so close as to be merged together according to modularity maximization. Figure 3 shows the distribution of the number of neighbors where we can notice that most communities have 1 to 5 neighbors.

In our scenario, neighbors are useful because they contain information that is relevant to the main cluster in a different way that can be useful discovering relationships and serendipity. Table 1 shows a sample of clusters with their respective neighbors ranked by similarity. In the case of the `gucci` brand, the neighbors describe diverse types of products but also a competitor. In the `iphone` example, not only a competitor is mentioned but there is a neighbor that anticipates the next move (i.e., `where to buy`). The examples for `refrigerator` and `columbia` include companies that sell such products. Finally, the last example contains more diverse choices.

2.2 Packaging

As mentioned earlier, web clustering engines work as a meta search engine by issuing a query to an existing web search engine index and processing the returned titles and snippets from the search engine result page. We follow a similar approach by issuing a query to our set of pre-computed query clusters and returning the main cluster, its neighbors, and the queries per each cluster.

For packaging (a form of clustering) the cluster and its neighbors, we adopt a suffix-array as the main data structure, which is also common for this type of solutions [3], and incorporate a number of heuristics for traversing the array. A nice property of suffix arrays is that all the suffixes are close in proximity for fast lookup.

The algorithm (presented in listing 1) works in two phases as follows. We first *exploit* the retrieved cluster and its content and then *explore* the neighbors. In the first phase, given the user query, we locate all suffixes in the suffix array that are next to each other using binary search. This subset forms the base subset for traversing the suffix tree and selecting the corresponding queries (labels) as the first and most important cluster, the *base cluster*. Figure 4 shows the base cluster for "`dishwasher`" and how the tree is derived (i.e., indexes 29->28->27 produce "`18 inch dishwasher`"). The text "`18 inch dishwasher`" is selected as final because there is an equivalent query in the input data whereas entries like "`depot dishwashers`" or "`at lowes`" are not because they are likely to be part of bigger query string.

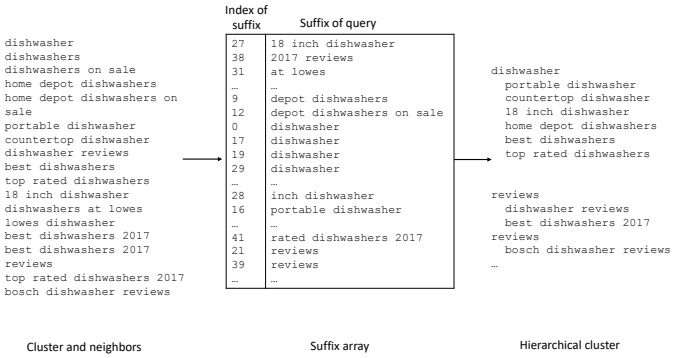


Figure 4: Example of how the suffix array is derived from the input data and how the base cluster is constructed by following the index. User query is {dishwasher}.

The base cluster is very relevant to the query but not sufficient for discovering information related to the topic. By computing term frequency (TF) for each n-gram of every suffix, we can identify high frequency terms that are candidates for exploration, that is, a second round of clustering. We iterate over a set of n-grams with high TF and produce another set of clusters. Finally, we merge the base clusters derived from the input query and the clusters produced by the n-grams with high TF into the final output. Compared to traditional clustering techniques like k-means, suffix arrays are very efficient for implementing clustering on small pieces of text.

Data: Query clusters computed by modularity
Input : User query q
Output : Clusters of queries relevant for query q

$qclusters \leftarrow$ find clusters and neighbors for q
/* parse input data, tokenization and stemming */
$sarray \leftarrow$ create suffix-array ($qclusters$)
/* exploitation */
$stree \leftarrow$ find query in $sarray$
$bcluster \leftarrow$ traverse $stree$ and extract labels
/* exploration */
$terms \leftarrow$ select terms from $sarray$ order by TF
$oclusters = \emptyset$
foreach t *in* $terms$ **do**
  $tc \leftarrow$ traverse $stree$ and extract labels
  $oclusters \leftarrow oclusters \cup tc$
end
$clusters \leftarrow$ merge($bcluster$, $oclusters$)
print $clusters$

Algorithm 1: Packaging clusters algorithm

Table 1: Sample of clusters and their respective neighbors ranked by similarity. Each cluster contains many queries.

gucci	iphone	refrigerators	columbia	dinner table
gucci glasses	iphone x	fridge	columbia clothing	round glass dining table
gucci bags	iphone 6s	home depot refrigerators	outdoor clothing for men	dining room sets
gucci sneakers	apple iphone	discount appliances	llbean sale	kitchen chairs
designer purses	iphone camera	refrigerator reviews	orvis men's clothing	farmhouse table
gucci handbags	where to buy	costco refrigerators	vantage apparel	small kitchen tables
chanel handbags	samsung phones	freezerless refrigerator	columbia jackets for men	amish tables
gold clutch	iphone review	kitchen appliance packages	coats	dining room furniture

3 RESULTS AND IMPLEMENTATION

The back-end pipeline runs on Cosmos, Microsoft's internal big data analysis platform. We use a data set of 24M queries from Bing categorized as commerce-related collected over a 1-month period. In our data set, there were 72M distinct links and the average sparseness of the query vector was 7.8 non-zero elements (clicked links per query). All these data is employed for constructing 3M query clusters using the technique described in Section 2.1. The top query within a cluster is indexed for search purposes.

At run-time, the computation works as follows. Once the user has entered a query, the system finds the most relevant query cluster and its neighbors ranked by similarity and returns a list of cluster labels and top queries from each cluster. This list is then clustered again using the method from Section 2.2 and presented as a Javascript tree explorer on a web browser where the user can explore the content. Each leaf entry triggers a query in a search engine allowing the user to examine the search results provided by the different views in the same user interface. Figures 5 and 6 show examples of our techniques in action.

The objective of the system is to group similar queries so the user can make sense of previous searches and intents when looking for a product. Instead of recommending related searches like in traditional web search engine results, we organize queries according to similar intents with the goal of providing the user the big picture when searching for products.

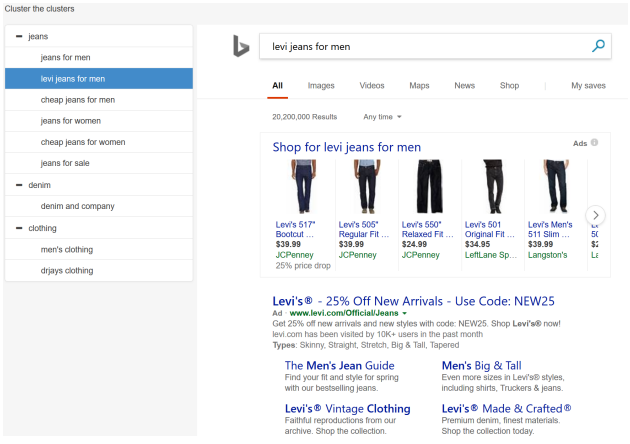


Figure 5: Search results for the query jeans.

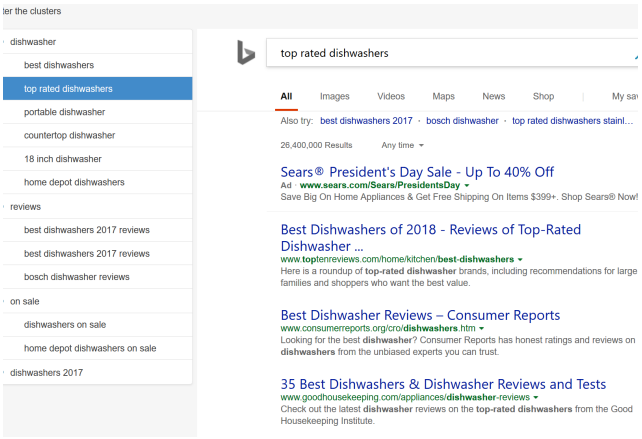


Figure 6: Search results for the dishwasher example.

4 CONCLUSION

We described a prototype that uses query logs and click data to cluster similar queries for grouping similar intents and how to cluster the output of such clusters to organize search results. The methods presented do scale and have been implemented in a working system. The demo is built only using queries instead of documents or surrogate pages like in previous systems. We also provided details on how the techniques work so they can be reproduced with other data sets. Future work includes the design of a user study for comparison against related queries.

REFERENCES

[1] Ricardo A. Baeza-Yates and Alessandro Tiberi. 2007. Extracting semantic relations from query logs. In *Proc. of SIGKDD*. 76–85.

[2] Ilaria Bordino, Mounia Lalmas, Yelena Mejova, and Olivier Van Laere. 2016. Beyond Entities: Promoting Explorative Search With Bundles. *Inf. Retr. Journal* 19, 5 (2016), 447–486.

[3] Claudio Carpineto, Stanislaw Osinski, Giovanni Romano, and Dawid Weiss. 2009. A survey of Web clustering engines. *ACM Comput. Surv.* 41, 3 (2009), 17:1–17:38.

[4] Marti A. Hearst. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press.

[5] Beibei Li, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. A Demo Search Engine for Products. In *Proc. of WWW*. 233–236.

[6] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[7] M. E. J. Newman. 2006. Modularity and community structure in networks. 103, 23 (2006), 8577–8582.

[8] Thibault Sellam, Martin Hentschel, Vasilis Kandylas, and Omar Alonso. 2016. e#: Sharper Expertise Detection from Microblogs. In *Proc. of EDBT*. 563–572.

[9] Daria Sorokina and Erick Cantu-Paz. 2016. Amazon Search: The Joy of Ranking Products. In *Proc. of SIGIR*. 459–460.

Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models

Anjan Goswami[**]
agoswami@ucdavis.edu
University of California Davis

Prasant Mohapatra
pmohapatra@ucdavis.edu
University of California Davis

Chengxiang Zhai[†]
czhai@illinois.edu
University of Illinois
Urbana-Champaign

ABSTRACT

The demand generation and assortment planning are two critical components of running a retail business. Traditionally, retail companies use the historical sales data for modeling and optimization of assortment selection, and they use a marketing strategy for demand generation. However, today, most retail businesses have e-commerce sites with rapidly growing online sales. An e-commerce site typically has to maintain a large amount of digitized product data, and it also keeps a vast amount of historical customer interaction data that includes search, browse, click, purchase and many other different interactions. In this paper, we show how this digitized product data and the historical search logs can be used in understanding and quantifying the gap between the supply and demand side of a retail market. This gap helps in making an effective strategy for both demand generation and assortment selection. We construct topic models of the historical search queries and the digitized product data from the catalog. We use the former to model the customer demand and the later to model the supply side of the retail business. We then create a tool to visualize the topic models to understand the differences between the supply and demand side. We also quantify the supply and demand gap by defining a metric based on Kullback-Leibler (KL) divergence of topic distributions of queries and the products. The quantification helps us identifying the topics related to excess or less demand and thereby in designing effective strategies for demand generation and assortment selection. Application of this work by e-Commerce retailers can result in the development of product innovations that can be utilized to achieve economic equilibrium. We can identify the excess demand and can provide insight to the teams responsible for improving assortment and catalog quality. Similarly, we can also identify excess supply and can provide that intelligence to the teams responsible for demand generation. Tools of this nature can be developed to systematically drive efficiency in achieving better economic gains for the entire e-commerce engine. We conduct several experiments collecting data from Walmart.com to validate the effectiveness of our approach.

[*]The author worked for Walmart Lab. when he conducted this research.
[†]The author was an advisor for R&D teams at Walmart Lab when he worked on this problem.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316605

CCS CONCEPTS

• **Information systems** → *Business intelligence.*

KEYWORDS

E-commerce search; Information retrieval; Business Analytics; Topic Models; Marketplace economics

ACM Reference Format:
Anjan Goswami, Prasant Mohapatra, and Chengxiang Zhai. 2019. Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316605

1 INTRODUCTION

Demand generation and assortment selection are two critical components of a retail business. Demand generation is essential for growing the business by acquiring customers interested in purchasing from the retail shop while optimal assortment selection contributes to the maximization of sales and revenue. These two areas always attracted substantial investment from retail businesses. The demand generation is typically achieved by employing effective marketing strategies. Often, this involves making statistical models of the customer segments who may be interested in some of the products of a retailer and reaching out to those people. The assortment selection algorithms use actual sales and product similarity data for optimizing the selection of assortments for maximizing the sales or revenue. However, today, most retailers also posses an e-commerce business, and the market share of sales from the e-commerce business has been rapidly increasing compared to the brick and mortar stores throughout the last decade. The e-commerce sites typically maintain a vast amount of digitized product data that includes product title, description, it's pictures, brand, price and many other attributes. E-commerce websites log user information and actions, require identification for purchase, and provide mechanisms by which users can explicitly state their demand. This is most notable in e-Commerce search where a user can query for the products of interest. Hence, by understanding the user queries, e-commerce companies can gain insight into the demand. Moreover, this demand data can then be coupled with product supply data to further understand the dynamics of supply and demand on e-commerce sites.

The preponderance of such data allows us to utilize statistical methods in understanding the real demand and supply for the products and develop plans to achieve an economic equilibrium as well as guide investment decisions on search engine optimization

and marketing (demand generation techniques), assortment choices, and technology.

In the case of large-scale e-commerce businesses that operate highly trafficked sites, the volume and variety of queries are well beyond meaningful human interpretation. To utilize this data and to make supply and demand decisions, e-commerce retailers require a scalable way to understand the broader themes on both sides and to be able to connect them. Topic models are unsupervised machine learning algorithms for mining data in large corpora that can uncover the underlying semantic structure of large data sets. They have been successfully applied to various types of data such as text, images, and biological data amongst others. In this paper, we present a novel application of topic models to systematically identify the supply and demand gap of an e-commerce engine from the textual contents of the query and the product data. We then correlate the gap with the revenue to quantify this. Identifying and quantifying the supply and demand gap is a practical problem for an e-commerce business. Using this, we can determine the cluster of queries that represent unmet demand on the site and the items that are rarely queried for and may need to be replaced. In this paper, we show that computing a distance between the topic models of a representative sample of queries on the topic space of the items for the e-commerce site can be a systematic way to obtain insights for both demand generation and the assortment planning. We evaluate our algorithm with a simulation study conducted using an extensive data set consisting of the historical search logs and the product catalog obtained from Walmart.com site.

Our paper makes the following key contributions:

- We address the problem of gaining insight about the supply and demand side gap for an e-commerce site from its search queries data and the product catalog using textual information.
- We propose an algorithmic framework that constructs topic models from query and items data and quantifies the gap between supply and demand.
- We provide an understanding of the economic dynamics of an e-commerce site so that areas most pertinent to a disequilibrium in supply and demand side can be identified using visualization and quantification.

2 BACKGROUND

In this paper, we extend that idea to show an emerging application of using topic models in providing insights about the potential directions of demand generation and assortment planning.

2.1 Topic models

Topic models are probabilistic generative models of the documents in a corpus. It assumes that each document consists of a mixture of topics that are shared across the corpus and each topic is a distribution over words. Once a model is learned, a document can be represented by a vector of topic probabilities. The strength of topic modeling is that it is an unsupervised algorithm and it requires only the corpora as input to compute the semantic relationship of the documents in the form of topics. One of the popular approaches of topic modeling is latent Dirichlet allocation (LDA) [7] that uses a Bayesian generative model. LDA algorithm has been proved to be successful in various applications beyond information retrieval and text mining including collaborative filtering, computer vision, and bioinformatics [4, 7, 16, 17]. The LDA algorithm's parameters are estimated using either sampling-based algorithms which attempt to approximate the posterior distribution from an empirical distribution or variation-based algorithms which try to find a distribution of best fit to the posterior by using a parameterized family of distributions over it [5]. Typical applications of LDA primarily use sampling-based algorithms, the most popular of which is Gibbs sampling which has been initially demonstrated by Griffiths and Steyver [12]. In this paper, we use the LDA algorithm to construct topic models for the set of queries in a fixed period and the set of documents from the product catalog that are in the inventory during that time. The queries and product titles are short text, and a direct application of LDA will not work. Hence, we first construct clusters of similar queries and items. Researchers [19] used similar clustering techniques for building topic models for short text.

2.2 Retail studies

Demand generation in retail is related to investing in marketing and advertising to attract potential customers to its product offerings. This involves traditional marketing [2], communicating promotions and deals to a potential audience [1], advertising the products in search engines or social networks [13, 21] etc. The assortment planning also has a rich literature in retailing [15]. It depends on many factors and often the process can be very complicated because several aspects of bringing a new product lines need to be considered [18]. It has been understood in retailing literature that it makes sense to get the insight about assortment planning from the consumer's search of products [8]. In this paper, we aim to obtain insight from actual e-commerce search data. Note that our algorithm intends to help in identifying some areas which are worth investigating for demand generation and assortment planning and hence it can be viewed as a complementary tool for the existing research in retailing in this area.

2.3 Distance between distributions

This is a classic area in statistics [3]. We use Kullback-Leibler divergence (KL-divergence) which is a popular such distance function in information retrieval [22] and has been in the past used to measure the quality of topic models [14].

3 TOPIC MODELS FOR E-COMMERCE SUPPLY AND DEMAND GAP

In this paper, we are interested in measuring the supply and demand gap of an e-commerce engine to understand (a) marketing needs to generate demand (b) assortment selection for unmet demand. Our basic idea is straightforward. We construct topic models from the textual content of the queries and the items. The topic models capture the underlying topic structure of the implicit economics of the e-commerce engine and define a distribution of topics in each corpus. We quantify the difference between these two topic distributions with a statistical distance function which measures the extent of the gap or disequilibrium between the supply and demand of an e-commerce engine. We then also use the existing click data to understand the business implication of the gap in terms of a click

based engagement metric. Note that, we use only the click data but it is simple to extend this using sales or revenue data and we can then also obtain the expected impact on sales and revenue based business metrics. We visualize the differential word clouds from the top words of topic models from the supply and the demand side to obtain a visual insight about the supply and demand gap.

3.1 LDA on Search Query and Items data

The demand side of an e-Commerce platform can be represented by the set of all queries that are submitted to the search engine in a fixed period of time. The supply side can similarly be represented by the set of all items in the inventory during the same period of time. Each item can be associated with some textual data that typically includes the title, and the description. This textual data can be used for topic modeling. The title is expected to convey a good summary of any items. The description can have a lot of information and can be somewhat noisy. In our experiments, we combine the title and the description data that is generally available for most of the items in a catalog for a retail company. We henceforth mention the combination of this textual data from the items as documents. We compute two topic models, the first one is from the set of queries that represents the topics on the demand side and the second one is from the items that represent the supply side.

We intend to make these topic models per category. We thus first start preprocessing the corpus removing the stop words, punctuation, and other semantic elements that do not have meaning in the context of topic models. The documents are tokenized and tokens with low term frequency (TF) and low inverse document frequency (IDF) [20] are removed from the set. Then, we construct a fixed vocabulary of tokens from a category that we use for topic modeling. Let's call this vocabulary V. One of the challenges of applying topic models for the query data is that queries are very short snippets of text and we cannot build topic models only from such short texts. We thus construct a mechanism to cluster similar queries by grouping them based on a common item that is clicked when it is shown in search results for these queries. We filter out item-query document pairs below certain traffic and click threshold. This gives us sets of small clusters of similar queries. We do not have a similar challenge to apply topic modeling for the items data since typically items come with a document with product description and attributes along with the product titles. We use the topic modeling based on latent Dirichlet allocation (LDA) for our problem. It works first by assuming a fixed number of topics K and then by drawing the topics from a distribution over the fixed vocabulary V. This distribution is assumed to be Dirichlet. After this, we generate a proportion of topics per document using another K dimensional Dirichlet distribution. We then further generate a topic assignment per word per document using a multinomial distribution and also generate the words for the topics selected using yet another multinomial distribution. The central computational problem for LDA is the estimation of the joint posterior probability of hidden topics given the observation of the documents. We use a classic variant of LDA algorithm described in a paper by Blei et al [6] where this problem is solved using Gibb's sampling and variational inference. Readers can find the details of the LDA algorithm in the paper by Blei et al [6].

We discuss the computation of the supply-demand gap in section 3.2.

3.2 Quantifying Supply-Demand Gap

In this section, we provide an algorithm to quantify the supply-demand gap. We define two distributions p_d and p_s as the distribution over words for demand and supply respectively. The distance function is defined as $dist(p_s, p_d)$. We use Kullback-Leibler divergence to measure the similarity between the two probability density distributions:

$$D_F(p, q) = \sum p(i) \log \frac{p(i)}{q(i)}$$

KL divergence is non-negative and is 0 only when the two distributions are themselves equal. KL divergence is, however, asymmetrical. In order to make it symmetric, we define our distance function $dist(p_s, p_d)$ as:

$$dist(p_s, p_d) = \frac{D_F(p_s, p_d) + D_F(p_s, p_d)}{2}$$

4 EXPERIMENTAL EVALUATION

We use the data from Walmart's e-commerce site which is a high traffic website and one of the largest e-commerce site in the United States. We evaluate our algorithm using six month's of search query data and the product catalog data available during that time from Walmart's e-commerce platform. We create two data sets. In one data set, we have the query, items that are shown for the query, and the action taken by the customers such as clicks, sales. We also keep the price data to compute revenue. On the other hand, the product data consists of title, description and set of attributes of the product.

We use the R package "lda" [9] that uses a Gibbs sampling-based method for inference. We generated a 100-topic LDA model fit to the search query and the item data.

We show the results of four experiments.
(1) First we show visualizations of topics from queries to explain the effectiveness of our algorithm for applying LDA to the set of queries.
(2) We then show some top words per category to show the effectiveness of applying LDA to the items.
(3) We show the word clouds from query and the item side. These word clouds can be excellent tool to provide visual insight on the supply and demand gap in an e-commerce site.
(4) We also conduct an experiment to show the effectiveness of our KL-divergence based statistical distance between topic distribution to capture the relationship with engagement. Intuitively, we expect to observe that if the distance is more then there will be less clicks.

4.1 Visualization of Search Query Topics

In table 1, we show five top words from seven topics from the LDA models of the query clusters. This can provide an idea about effectiveness of our application of LDA to our search query (cluster) data.

Topics	word 1	word 2	word 3	word 4	word 5
1	frozen	disney	doll	elsa	queen
18	curtain	decorative	drape	sheer	panel
23	kayak	boat	fishing	inflatable	rod
31	sofa	futon	couch	leather	sectional
53	bathroom	shower	bath	towel	mainstays
64	stroller	carseat	graco	britax	combo
97	bed	full	twin	frame	platform

Table 1: Seven topics from a 100-topic LDA model fit to search log. A simple inspection of the words tells us the following about the topics: topic 1 is about frozen, the movie, topic 18 is about curtains and drapes, topic 23 is about fishing and kayaks, topic 31 is about sofas, topic 53 is about accessories of bathing, topic 64 is about carseats, and topic 97 is about beds.

4.2 Visualization of Items Related Topics:

In figure 1 we show six plots where the X-axis shows the uniqueness of the words which is the inverse document frequency and the Y-axis shows the term frequencies. The terms on the upper right corner are high in uniqueness and also high in frequency. We select some of the top words from some of the topics generated from the product catalog dataset. It is easy to see that each topic represents a specific product theme.

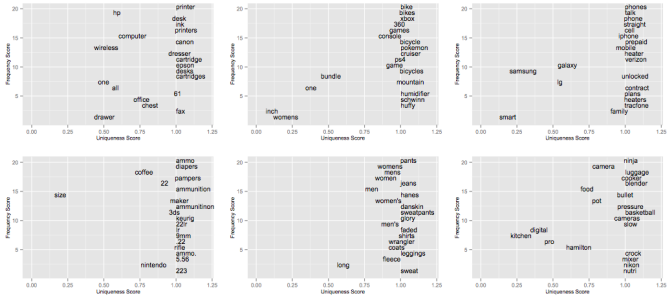


Figure 1: Frequency vs. uniqueness plots of words for select topics. Upper left plot can be considered about Samsung TV, the upper middle plot may have a few different products from furniture, upper right one is about smartphones, lower right one is about carseat, lower middle one is about laptops.

4.3 Visualization of Supply and Demand Gap

Now, we show some visualizations of the supply and demand sides topics using word clouds. This allows us to identify areas where there is excess supply and demand. In figure 2 we show a word cloud of the demand and supply of a given topic side by side. In order to obtain the visualizations, we weigh the demand side words by query traffic obtained for search log data and for the supply side, we weigh the item titles by the item count in the inventory from the product catalog. From visual inspection, we can identify that some large words like “women" and “shoe" are common in both our demand and supply sides. This indicates that customers may be searching for “Women’s’ Shoe” and the inventory also have adequate stock for Shoes.

Figure 2: Word clouds on demand and supply for one topic

To identify areas of opportunity, we remove these common words and regenerate the word cloud visualization as shown in Figure 3. The resulting figure is much more informative. We observe now demand for which there is not comparable supply and vice versa. For this topic, we can clearly see that there is excess demand for “jansport" bags for which our supply appears deficient. Our supply side word cloud now consists of tokens of item titles that receive lacking consumer interest such as “textil" (eg. textile) and “fuchsia" for which there is little demand.By generating similar word cloud visualizations of the supply-demand gap for various topics, we can systematically identify areas for improvement in an e-commerce engine.

4.4 Validation of the Supply Demand Gap Metric

In this study, we compute our distance function $dist(p_s, p_d)$ to quantify the supply-demand gap. In order to understand its relationship with an engagement metric, we artificially remove some products or queries from the dataset and generate the topic models and compute the distance. We see, in figure 4, that the clicks are correlated to the gaps between supply and demand. The larger the gap, we see a larger drop in clicks. This informs us that we can use such divergence measures as a metric to understand the degree of supply and demand matching from the text data in an e-commerce site. This will allow us to understand the economics of supply and demand without the historical data. The method also can be fully automated. Hence, this can be a very powerful metric in e-commerce to quickly conduct a first level automated health check of the e-commerce site considering the supply and demand gap.

Figure 3: Word clouds of relative complements

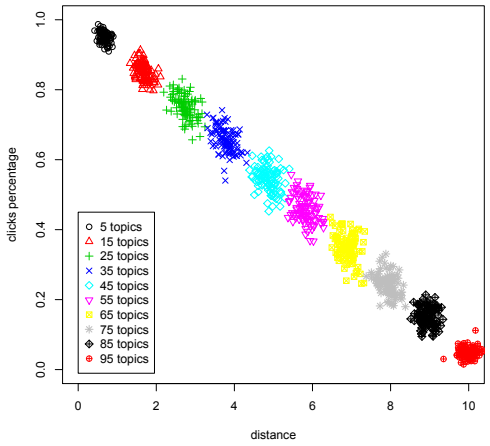


Figure 4: Relationship between supply-demand distance and clicks, we randomly remove 5 to 95 topics from the inventory and draw the correlation plot with clicks. Here, it shows clicks increase monotonically with less statistical distance between the query item topic distributions.

5 CONCLUSION

In this paper, we develop a topic models based approach to obtain insight about demand generation and assortment planning of an e-commerce site. We also define a measure of supply and demand gap on an e-commerce site and show that this measure is monotonically decreasing with the click-based engagement metric. Hence, if this measure indicates a large gap, then it is not healthy for an e-commerce site. The technique is based on topic models, is unsupervised, scalable, and can be very powerful in determining the state of supply and demand in a large e-commerce site. The visualizations also can help in obtaining visual insights on the demand and supply side. The method can be used as a powerful unsupervised technique to understand the supply and demand gap in an e-commerce site. Future work can be done to utilize the existing tree structure of categories to analyze supply-demand gaps as they pertain to a retailer's organizational and product hierarchy. Such work can result in a more granular and operationally effective application of topic models to increase economic output for the e-commerce business. Additionally, we can explore other clustering strategies beyond clicks to handle short text documents to improve the topic model such as different metrics like add-to-cart or conversion data and using Bipartite Spectral Graph Partitioning [10]. Additionally, a similar methodology can be used by constructing topic models using any deep learning and word embedding models [11]. We believe that our paper can pave the way for exploring with similar advanced natural language processing based tools to identify demand generation and assortment choices from the vast amount of search log data for e-commerce sites.

ACKNOWLEDGMENTS

We acknowledge Walmart Inc. for providing us the permission to publish the paper. We also acknowledge Angela Jiang and Wei Han for their contribution to this work when they worked at Walmart Lab. An earlier version of this work coauthored with Angela Jiang and Wei Han was presented in 2015 INFORMS data mining and analytics workshop.

REFERENCES

[1] Kusum L Ailawadi, Jonathan P Beauchamp, Naveen Donthu, Dinesh K Gauri, and Venkatesh Shankar. 2009. Communication and promotion decisions in retailing: a review and directions for future research. *Journal of retailing* 85, 1 (2009), 42–55.
[2] Gary Armstrong, Philip Kotler, Michael Harker, and Ross Brennan. 2015. *Marketing: an introduction.* Pearson Education.
[3] Michéle Basseville. 2013. Divergence measures for statistical data processing?An annotated bibliography. *Signal Processing* 93, 4 (2013), 621–633.
[4] Indrajit Bhattacharya and Lise Getoor. 2006. A latent dirichlet model for unsupervised entity resolution. In *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 47–58.
[5] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826
[6] David M Blei, Michael I Jordan, et al. 2006. Variational inference for Dirichlet process mixtures. *Bayesian analysis* 1, 1 (2006), 121–143.
[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937
[8] Gérard P Cachon, Christian Terwiesch, and Yi Xu. 2005. Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management* 7, 4 (2005), 330–346.
[9] Jonathan Chang. 2011. R package 'lda'. http://cran.r-project.org/web/packages/lda/
[10] Inderjit S. Dhillon. 2001. Co-clustering documents and words using Bipartite Spectral Graph Partitioning.
[11] Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57 (2016), 345–420.
[12] Thomas L. Griffiths, Mark Steyvers, Thomas L. Griffiths, and Mark Steyvers. 2004. Colloquium Finding scientific topics.
[13] Lisa Harris and Charles Dennis. 2011. Engaging customers on Facebook: Challenges for e-retailers. *Journal of Consumer Behaviour* 10, 6 (2011), 338–346.
[14] Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics (SOMA'10)*. ACM, New York, NY, USA, 80–88. https://doi.org/10.1145/1964858.1964870
[15] A Gürhan Kök, Marshall L Fisher, and Ramnath Vaidyanathan. 2008. Assortment planning: Review of literature and industry practice. In *Retail supply chain management*. Springer, 99–153.

[16] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 61–68.

[17] Marie Lienou, Henri Maitre, and Mihai Datcu. 2010. Semantic annotation of satellite images using latent Dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters* 7, 1 (2010), 28–32.

[18] Murali K Mantrala, Michael Levy, Barbara E Kahn, Edward J Fox, Peter Gaidarev, Bill Dankworth, and Denish Shah. 2009. Why is assortment planning so difficult for retailers? A framework and research agenda. *Journal of Retailing* 85, 1 (2009), 71–83.

[19] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and Sparse Text Topic Modeling via Self-Aggregation.. In *IJCAI*. 2270–2276.

[20] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. 133–142.

[21] Sha Yang and Anindya Ghose. 2010. Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science* 29, 4 (2010), 602–623.

[22] Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 403–410.

ProductQnA: Answering User Questions on E-Commerce Product Pages

Ashish Kulkarni*
India Machine Learning, Amazon
kulkashi@amazon.com

Kartik Mehta*
India Machine Learning, Amazon
kartim@amazon.com

Shweta Garg*
India Machine Learning, Amazon
shwegarg@amazon.com

Vidit Bansal*
India Machine Learning, Amazon
bansalv@amazon.com

Nikhil Rasiwasia
India Machine Learning, Amazon
rasiwasi@amazon.com

Srinivasan H Sengamedu
India Machine Learning, Amazon
sengamed@amazon.com

ABSTRACT

Product pages on e-commerce websites often overwhelm their customers with a wealth of data, making discovery of relevant information a challenge. Motivated by this, here, we present a novel framework to answer both *factoid* and *non-factoid* user questions on product pages. We propose several question-answer matching models leveraging both deep learned distributional semantics and semantics imposed by a structured resource like a domain specific ontology. The proposed framework supports the use of a combination of these models and we show, through empirical evaluation, that a cascade of these models does much better in meeting the high precision requirements of such a question-answering system. Evaluation on user asked questions shows that the proposed system achieves 66% higher precision[1] as compared to IDF-weighted average of word vectors baseline [1].

CCS CONCEPTS

• **Information systems** → **Question answering**; • **Applied computing** → *Online shopping*.

KEYWORDS

question answering; deep learning; chatbot; e-commerce

ACM Reference Format:
Ashish Kulkarni, Kartik Mehta, Shweta Garg, Vidit Bansal, Nikhil Rasiwasia, and Srinivasan H Sengamedu. 2019. ProductQnA: Answering User Questions on E-Commerce Product Pages. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316597

1 INTRODUCTION

Online e-commerce systems play a vital role in connecting product sellers and end consumers at scale. However, consumers often struggle to navigate through the millions of products on offer and therefore, the success of these systems relies on their ability to seamlessly support customers in their product discovery and research. This has motivated a lot of work in the areas of product search, recommendation, information extraction, summarization, and recently, automatic question answering [17, 20] and chatbots [22]. In this work, we are concerned with the specific problem of answering customer questions on e-commerce product pages. Product detail pages often contain a wealth of information contributed by both sellers (product title, description, features, *etc.*) and customers (reviews, community question-answers, *etc.*). However, in their effort to offer the most comprehensive product information, the amount of data on these pages has grown so much, that for a top selling product, the detail page typically spans over six to eight thousand words, filling up around 15 A4 sheets. Customers also face an increased complexity in product evaluation due to variations ("size" *vs.* "dimension") and implicit references to product features (e.g. for title "20.1 *MP* Point and Shoot Camera *Black*", 20.1 *MP* refers to resolution and *Black* refers to color attribute). On small form factor devices like mobile, customers might benefit from a system that answers their product-related questions without having to browse through the page.

Building such a question-answering system poses some interesting challenges.

Question intent: In addition to product feature-related questions (like, "size" or "resolution"), customers could ask other *factoid* questions like "what's in the box?", "does this work with canon?" or *non-factoid* questions like "is this worth the money?" Understanding question intent is key to generating an appropriate response.

Product attribute-value: The system should account for explicit and implicit references to product attributes and their values in both questions and candidate answer lines.

Semantic matching: Customers often use text variations (*eg.*"anti-shake" to refer to "image stabilization"), thus necessitating semantic matching of question and answer lines.

High precision: Providing incorrect answers would lead to a marred customer experience and add to their frustration.

Lack of training data: Unlike question answering systems for open domain, domain specific systems suffer from scarcity of training data and other resources like structured knowledge bases.

Addressing these challenges for domain specific question answering systems is the primary focus of this work. We believe that building such a system would involve an interplay of different components for identifying question intent, attribute name-value

*These authors made equal contribution
[1]Evaluated at fixed coverage, where coverage is the number of questions that receive an answer. We cannot reveal the exact coverage number due to confidentiality

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316597

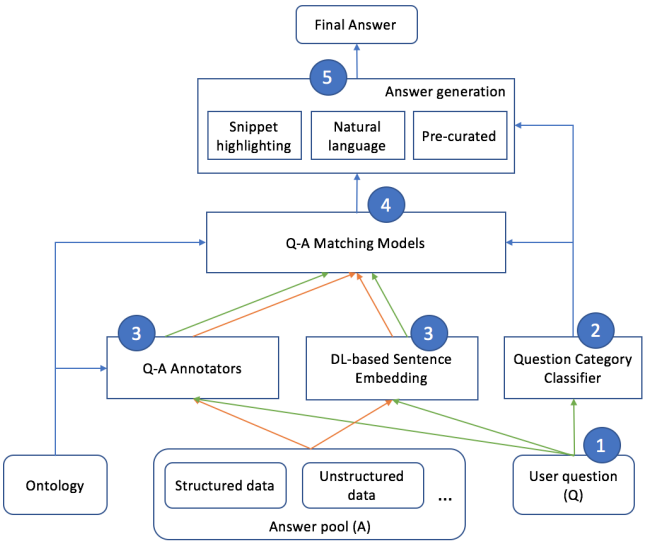


Figure 1: Framework for question-answering leveraging structured and distributed semantics. ① Framework receives a user question; ② Question category classifier classifies the question into one of the predefined categories; ③ Question and answer sentences are processed to generate their ontology-based annotations and deep learning-based embeddings; ④ Matching models rank the answer sentences for their relevance to the question; ⑤ Answer generation component generates the final answer based on the ranked answer sentences.

annotation-based on a structured knowledge base, semantic matching of question and answer lines, and final answer generation. We present a generic framework for in-domain question answering. The framework allows for a graceful combination of deep learning-based distributed semantics and semantics imposed by a structured resource like a domain ontology. Along with a question classifier to identify intent, the proposed system caters to the high precision requirement for a great customer experience. We present a detailed evaluation of different components of the framework and an ablation study underlining their contribution to the system performance.

2 RELATED WORK

The body of work closest to the proposed framework comes from the field of question answering for e-commerce. Yan *et al.* [22] recently presented a task-oriented dialog system that leverages an in-domain knowledge base, search logs and community sites to assist users for online shopping. Distinct from them, SuperAgent [3] takes advantage of in-page product descriptions and user-generated content to answer user questions for a product. While we are also concerned with in-page question answering, we present a more generic solution covering aspects of question understanding, question-answer representation and matching and answer generation. We support the efficacy of the proposed framework via a detailed empirical study.

Contribution of question answering and reading comprehension datasets, notably, TREC [18] and recently, SQUAD [13] and MS MARCO [11] has led to a lot of work in the area of open-domain question answering of factoid questions from a given document collection. Some of the earlier systems [14] made use of text and entity-level surface patterns as clues to right answers. Realizing that these approaches suffered from low recall and did not capture long-distance dependencies, some of the subsequent research extended these with other statistical signals from the corpus [15] or more complex patterns based on deep linguistics [12]. Other approaches based on hand crafted syntactic features [8] have also been explored. Although we are also concerned with answering user questions from a given passage of text, the domain of interest is limited (to e-commerce products, for instance), making it difficult to leverage existing language resources and knowledge bases in the open domain.

With deep learning gaining in popularity, there's a recent body of work in question answering that leverages dense representation of sentences composed from neural word embeddings [10]. Several sentence embedding approaches have emerged based on simple word vector averaging [21] or those leveraging the structure and sequence of words in a sentence using RNN, LSTM or CNN-based [6] architectures. When applied to the question answering task, some of the existing work is based on the semantic similarity of a question and a potential answer in a jointly learned embedding space [9], while others employ a classification or learning-to-rank approach over joint question-answer feature vectors [19]. While the proposed embedding models are inspired from some of the aforementioned approaches, we differ from them in that we complement the distributional semantics learned from these models with the structured semantics imposed by an ontology and combine these in a generic question answering framework. We show that a question answer matching model based on a combination of these features achieves much better results on an in-domain question answering task.

3 PRODUCTQNA FRAMEWORK

Figure 1 gives an overview of proposed ProductQnA (PQnA) framework. We are given a question q and a pool of candidate answer lines $\mathcal{A} = \{a_1, \ldots, a_n\}$. We then pose question answering as a ranking problem, where, the candidate answer lines are ranked based on their relevance to the question q and top-k answers $a'_1, \ldots, a'_k$ ($a'_i \in \mathcal{A}$) are selected for final answer generation if their relevance $s(a'_i)$ exceeds some threshold t. It is possible that none of the answer lines get selected if they all fail to meet the threshold.

We describe the ranking (or question-answer matching) models in more detail in the following sections. The matching models in the proposed question-answering framework (refer to Figure 1) are further aided by several other components which we also describe in detail below.

3.1 Ontology

An *ontology* describes the entity types in a domain and their interrelationships. We built an ontology for a large product category, where the entity types comprise products (camera, lens, tripod *etc.*), their attributes (dimension, resolution, *etc.*) and attribute values (20.1 MP, Black *etc.*) and the relationships capture their semantic relatedness, for instance, baby_monitor $\xrightarrow{\text{isA}}$ camera, security_camera $\xrightarrow{\text{hasA}}$

night_vision, resolution $\xrightarrow{\text{hasValue}}$ resolution_value. We bootstrap the ontology from existing in-domain knowledge bases and gazetteers (list of *colors*, *brands etc.*) and further augment it with entities extracted from semi-structured and unstructured corpus of product pages. Product attributes and their values often appear as feature bullets displayed in a tabular fashion on product pages. We exploit such structure on product pages to extract these attributes and their values. We also extract frequently occurring noun phrases, from the unstructured text, which are manually audited and merged into the ontology using Protégé[2]. The ontology that we thus curated, consists of 570 entity types spanning product categories like digital cameras, security cameras, lenses, tripods, bags and cases, batteries, films and others.

3.2 Question-Answer Annotators

An annotator extracts semantics from text by identifying entity mentions (like, *anti-shake* or 20.1 *MP*) in raw text and linking them to their canonical entities (*image_stabilization* and *resolution_value*, respectively) in an ontology. We annotate user questions and candidate answer lines to generate *annotations*, which are triples $\langle e, s_{begin}, s_{end} \rangle$, where, e is an entity in the ontology and s_{begin} and s_{end} define the span of the entity mention in the raw text line. We use three types of annotators:

Regular expression-based: Attribute values (*e.g.* 20.1 MP or 10 GB) often have a well defined signature and could be extracted using a regular expression annotator.

Gazetteer-based: Lists of certain attribute values like color, camera brand *etc.* are often readily available. We leverage these to define gazetteer-based annotators for attributes *color_value*, *camera_brand_value* and others.

Machine learning models: In order to capture semantic variations ("how long does this battery last?" is a reference to *battery_life*), we manually label annotations for a subset of user questions, $Q_{labeled}$ and use a k-NN classifier to annotate an unseen user question q. As distance metric, we use the Jaccard similarity between q and the questions in $Q_{labeled}$.

A union of the outputs from these annotators is then used as the final set of annotations, Q_{annot}, for a question and A_{annot}, for a candidate answer.

3.3 Deep Learning based Sentence Embedding

While annotators provide ontology-based semantic features for a sentence, we also use deep learning-based sentence embeddings leveraging distributional semantics of words and their context. The question and answer embeddings thus obtained serve as another input to the question-answer matching models. The embedding architecture (refer Figure 2) is inspired from the Siamese neural network [4]. Given a sentence, tokenized into words, the network takes as input their word embeddings, typically initialized with embeddings pre-trained on large in-domain corpora. These are then composed together in the following layers, using a bag-of-words or word sequence approach, to obtain the final sentence embedding. For the question-answering task, we project the question and a candidate answer in a shared embedding space and the network parameters are trained to minimize a task-specific loss function.

[2]https://protege.stanford.edu/.

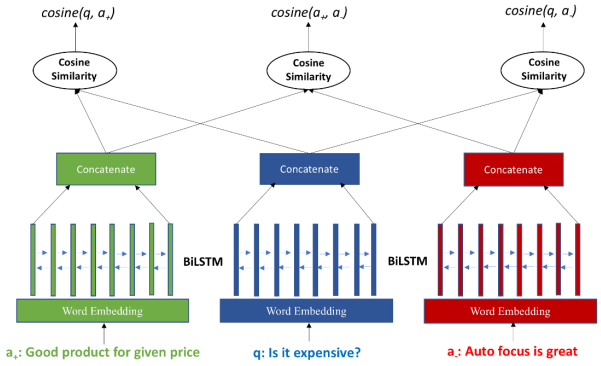


Figure 2: Model architecture for training deep learning-based sentence embedding. q is a question, a_+ is relevant answer to the question and a_- is any irrelevant statement.

We discuss the different sentence embedding approaches and loss functions below.

3.3.1 Sentence embedding using supervised word averaging: For a sentence $s = w_1 \dots w_n$, where, w_i is a word in s and $\boldsymbol{w_i} \in \mathbb{R}^d$ its embedding, the sentence embedding $\boldsymbol{l}$ is computed as: $\boldsymbol{l} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{w_i}$. We initialize word embeddings with random weights and learn them as part of supervised training. This simple approach of averaging word vectors has shown to give comparable performance to complex deep learning models such as LSTM for text classification [5] as well as text similarity problems [1, 21].

3.3.2 Sentence embedding using LSTM:. As against the bag-of-words approach above, LSTM takes the sequence of words into account. It produces a vector $\overrightarrow{\boldsymbol{l_t}}$ at each word w_t, from its word embedding $\boldsymbol{w_t}$ and that of its previous context $\boldsymbol{w_1} \dots \boldsymbol{w_{t-1}}$. In case of bi-LSTM, $\overleftarrow{\boldsymbol{l_t}}$ is similarly obtained by reversing the order of words in the sentence and taking into account $\boldsymbol{w_t}$ and its context $\boldsymbol{w_n} \dots \boldsymbol{w_{t+1}}$. The concatenation of output vector from each direction, $\overleftrightarrow{\boldsymbol{l}} = \overrightarrow{\boldsymbol{l_n}} || \overleftarrow{\boldsymbol{l_1}}$ is then used as the final sentence representation.

3.3.3 Loss functions: The embedding models discussed above are trained in a supervised manner, where the training data comprises triplets $\langle \boldsymbol{q}, \boldsymbol{a_+}, \boldsymbol{a_-} \rangle$ of embeddings of question, correct answer and an incorrect answer respectively. The training aims to minimize a task-specific loss function which we discuss next.

Weighted Log loss is defined in [7] as: $L_l = -\log p(\boldsymbol{q}, \boldsymbol{a_+}) - \eta \log(1 - p(\boldsymbol{q}, \boldsymbol{a_-}))$ where, $p(\boldsymbol{u}, \boldsymbol{v}) = 1/(1+\exp(-\boldsymbol{u}^T \boldsymbol{v}))$ and $0 < \eta \leq 1$ dampens highly representative negative samples in the training data. We use $\eta = 1$ in the experiments as we have balanced number of negative and positive samples.

Siamese Hinge loss is commonly used for *Siamese architectures* [9] and is defined as: $L_s = \max\{0, M - cosine(\boldsymbol{q}, \boldsymbol{a_+}) + cosine(\boldsymbol{q}, \boldsymbol{a_-})\}$, where M is the margin.

Triplet Hinge loss: We propose a stricter version of the above loss that additionally penalizes the similarity of $\boldsymbol{a_+}$ and $\boldsymbol{a_-}$. Also, inspired from [16], we use different margin for the three components of the loss. In our experiments, this loss function has been found to achieve better results than siamese hinge loss, as we discuss in

more detail in Section 5.1.

$$L_3 = \max\{0, M_1 - \cos(\boldsymbol{q}, \boldsymbol{a}_+)\} + \max\{0, \cos(\boldsymbol{q}, \boldsymbol{a}_-) - M_2\} + \max\{0, \cos(\boldsymbol{a}_+, \boldsymbol{a}_-) - M_3\} \quad (1)$$

3.4 Question-Answer Matching Model

The question-answer matching model receives as input the question and answer feature representations from the annotators and deep learning-based embedding models and generates a final list of answers. We use the following matching models.

Similarity-based ranking model: Given the question embedding $\boldsymbol{q}$ and answer embeddings $\{\boldsymbol{a_1}, \ldots, \boldsymbol{a_n}\}$, the similarity-based ranking model f_{deep} ranks the answers based on their cosine similarity $cos(\boldsymbol{q}, \boldsymbol{a_i})$ to the question in the shared embedding space. A ranked list of answers, with similarity score exceeding a threshold t, is generated as the output.

Annotation-based classification model: Let Q_{annot} and A_{annot} be the set of annotations for a question and a candidate answer respectively. The annotation-based classification model f_{annot} is a binary classifier that returns 1 if any entity $e_q \in Q_{annot}$ subsumes an entity $e_a \in A_{annot}$ and 0 otherwise. An entity e_i is said to subsume an entity e_j if at least one of these assertions holds true in the ontology: $e_i = e_j$, $e_j \xrightarrow{\text{isA}} e_i$, $e_i \xrightarrow{\text{hasA}} e_j$ or $e_i \xrightarrow{\text{hasValue}} e_j$.

Ensemble matching model: One could define an ensemble matching model combining the semantic signals from ontology-based annotations and deep learning-based embedding models. Here, we use a cascade of models, where, the candidate answers are first ranked based on f_{deep} and subsequently filtered by f_{annot} to generate a final list of top-k answers.

3.5 Question Category Classifier

Customer questions might span multiple categories (refer to Table 1). Identifying these might help in generating an appropriate response to the question. For instance, one could use question category as an additional feature to the matching models or have separate models based on question categories. Also, in order to maintain the high precision requirement, one might choose not to answer certain categories (e.g. *other*, where, often answer is not available on the page). Certain categories (*"greetings"*, *"shipping_delivery"*, *"warranty"*, *"returns_funds"*,*"used_refurbished"*) have limited surface forms and can be answered with precurated response. We term these categories as *stock categories* and the rest as *non-stock categories*.

Building such a question classifier poses multiple challenges: (1) class ambiguity (e.g. "how expensive is this camera compared to others" question is ambiguous with *price* and *related_product* as candidate classes), (2) spelling mistakes (e.g. "what is *prise*", "what is *brnad*"), (3) complex surface forms (e.g. "does it take picture" is *specs*, but "does it make sound when it takes picture" is *others*) and (4) multiple sub-questions. Also, lack of sufficient training data adds to the complexity of this problem. In order to deal with these challenges, we use deep learning-based architecture. Formally, given a question q, we learn a function $f(\boldsymbol{q})$ that maps it to one of the question categories $\{c_1, \ldots, c_k\}$ as in Table 1. While there are several choices to model $f(\boldsymbol{q})$ (refer to Section 5.2 for an empirical comparison), we use a CNN model similar to the one used by Yoon *et al.* [6]. We propose two extensions to this architecture to make the classifier robust to spelling mistakes and generalize to unseen specs attributes.

Table 1: Question categories and their proportion in data

Question Category	Example	Proportion
specs	What is the weight?	34.3%
compatibility	Will this work with Nikon D300?	10.8%
ratings_and_reviews	What is the customer rating?	5.8%
whats_in_the_box	What comes with camera?	3.6%
returns_refunds	How can I return this package?	2.3%
shipping_delivery	Can I get it delivered to India?	1.6%
related_product	what speaker are people using with the camera	1.6%
warranty	Does it come with a warranty?	1.4%
used_refurbished	Is this a new camera or a refurbished one?	1.0%
greetings	Good evening	0.9%
price	How much does it cost?	0.7%
gibberish	abcd	0.4%
other	How do you access the video footage?	35.6%

Enriching classifier with subword information: We augment our CNN-based question classifier with character n-grams (subwords) [2]. The resulting model (*CNN+Subw*) is found to be robust to spelling mistakes.

Enriching classifier with f_{annot}: Gathering training data for all *specs* attributes and their surface forms is a challenging task. f_{annot} (introduced in Section 3.4) could be used to annotate questions with attribute tags in order to reduce the training data sparsity. For instance, "what is resolution" is annotated as "what is specs_tag". We then train a multi-channel CNN [6], where we use two different inputs (original question for first channel and annotated question for the other channel). We refer to this model as CNN+Subw+f_{annot} and present empirical evaluation in section 5.2.

4 SYSTEM ARCHITECTURE

Based on PQnA framework discussed above, we propose a question answering system. Users can ask questions about the product and the system provides instant answers from three different sources - (1) seller provided *product data* (2) *user reviews* and (3) *community Q&A* (*CQnA*). User questions and all the product detail page data from the three sources are subjected to the proposed PQnA framework to generate the top-3 answers. The question category classifier first classifies the question into one of the question categories. For questions belonging to one of *specs, ratings_and_reviews, compatibility*, and *price*, we then rank the sentences for their relevance to the question using the ranking models. As discussed in Section 3.4, we use a cascade of f_{deep} and f_{annot} as the ensemble matching model

Table 2: Qualitative examples for PQnA system.

User Question	System Answer
How big is this camera?	• dimensions : 3.82 x 4.65 x 5.12 inches. • the camera is big, almost the size of dslr.
Would this be good for beginners?	• Perfect for the first-time GoPro user, or as a second camera. • HERO Session is simple and easy to use.
Does it work under water?	• waterproof to 50ft/ 15m, freezeproof to 14 f/ -10 c, shockproof to 5.8ft, and dustproof. • it takes excellent pics in and out of the water.
Does it have anti-shake?	• image stabilization : yes • also compensates for the shift-type camera shake common in macro shooting
How does it compare with rebel SL1?	**CQnA Question:** What is the difference between this camera and the Rebel SL1? **CQnA Answer:** The SL1 is more compact and the specs are not as good as the T5i
Is it expensive?	For the price this camera is good value for money for an amateur photographer.
Is it compatible with canon lenses?	The EOS Rebel T6 camera is compatible with all Canon lenses in the EF and EF-S lineup.

for *product data* and f_{deep} alone for *user reviews* and *CQnA* data. We use a set of pre-curated answers for questions belonging to *greetings, shipping_delivery, warranty*, and *returns_refunds*. Currently, we do not provide an answer to *whats_in_the_box, related_product* and *other* categories. Table 2 shows examples retrieved from the system.

5 EVALUATION

We use a random sample of 1340 questions (Table 1 shows the distribution) to evaluate the system for coverage (fraction of questions for which we retrieve an answer) and precision (fraction of questions for which top retrieved answer is correct). For comparison, we use IDF-weighted average of word vectors (referred as *IDF-vector-average* hereinafter) which has been found to be a strong baseline for textual similarity tasks [1]. At a fixed coverage[3], our proposed system (as described in Section 4) achieves 66% higher precision as compared to *IDF-vector-average*. In this section, we describe experimental setup and evaluation for various components of the system.

[3]The coverage point is decided based on business requirement. We cannot disclose the absolute value due to confidentiality

Table 3: Comparison of architecture and loss functions. Precision numbers are relative to *IDF-vector-average*

Model Architecture	Loss function	Precision	Latency (in ms) [4]
Supervised Word Averaging	Log loss	-14%	**49**
	Siamese Hinge Loss	21%	
	Triplet Hinge Loss	24%	
LSTM	Log loss	-26%	568
	Siamese Hinge Loss	21%	
	Triplet Hinge Loss	**30%**	

5.1 Evaluation of Matching Models

CQnA provides a natural source of labeled data to train the deep learning-based matching model, f_{deep}. For each question and answer pair in *CQnA*, we generate training triplets by sampling five incorrect answers. Though *CQnA* data consists of rich surface forms, it suffers from some gaps (e.g. Answer lines are often lengthy and are not representative of snippets from product data). In order to alleviate these limitations, we augment *CQnA* data with semi-automatically generated data by leveraging the ontology and question-answer annotators. For each attribute in the ontology, we create a set of question templates (for "price", the question templates include "what is the cost?", *etc.*). We run the annotators on product data to obtain the positive answers for each attribute and use negative sampling to obtain corresponding negative answers. The labeled set thus generated by combining both the resources, comprises over 15M training triplets. For evaluation, we use 10% of *CQnA* questions and sample hundred incorrect answers for each question. For a given question, we score all candidate answers (one correct and hundred incorrect answers) and return the answer with highest score. For each architecture, we compare precision on these highest scored candidates (*P@1*) and report numbers relative to *IDF-vector-average* baseline.

Table 3 shows performance of log loss, siamese hinge loss and triplet hinge loss (Eqn 1) for both LSTM and supervised word averaging models compared to *IDF-vector-average* baseline. Siamese network based loss functions (siamese hinge loss and triplet-hinge loss) perform significantly better than logloss. The proposed triplet-hinge loss (30% improvement for LSTM) has been observed to perform slightly better than standard hinge loss function (21% improvement for LSTM). This is likely due to the stricter nature of triplet-hinge loss function. While both the model architectures show comparative performance (with LSTM model performing slightly better), the supervised word averaging model is 11 times faster than the LSTM model during evaluation time[4]. Due to comparative performance and better latency, we use the supervised word averaging model in rest of the evaluation.

5.2 Evaluation of Question Category Classifier

We train the question category classifier on a randomly sampled set of 7000 questions and use a validation set of 700 questions for tuning

[4]These latency numbers are averaged over scoring 10000 ⟨ query, answer ⟩ pairs and were done on the following configuration: Intel(R) Xeon(R) CPU E5-2665 0 2.40GHz 8 core with 148.84 GB memory

Table 4: Comparison of different architectures for Question Classifier. All numbers reported are relative to FastText.

Model Architecture	Multi class Accuracy	AUC (weighted)
Logistic Regression	-6.5%	-3.1%
FastText	0.0%	0.0%
LSTM	-0.1%	0.4%
CNN	1.2%	0.3%
CNN+Subw	5.5%	2.6%
CNN+Subw+f_{annot}	**8.4%**	**3.3%**

the model hyper-parameters. We evaluate multiple classification models—logistic regression with bag-of-words features, FastText[5], LSTM and CNN. For the FastText architecture, we choose *ngram size* and *number of epochs* based on cross validation. We use similar setting as Yoon *et al.* [6] for CNN model and a single hidden layer of 128 dimensions for LSTM. We implemented the CNN and LSTM models in Keras[6] and chose best epoch based on performance on validation set. We use set of 1340 questions (refer to Section 4) and multi-class accuracy and multi-class weighted AUC for evaluation and report numbers relative to FastText architecture.

The logistic regression classifier trained with bag-of-word features leads to a drop of 6.5% in multi-class accuracy (refer to Table 4) signifying the complexity of the problem. The CNN model achieves 1.2% improvement whereas no significant improvements are observed using LSTM architecture. CNN with subword embeddings achieves 5.5% improvements and leveraging f_{annot} further improves this to 8.4%.

6 SYSTEM ABLATION STUDY

We analyze the effect of different components on system metrics using the same set of 1340 questions used for evaluating the overall system. We compute precision (*P@1*) and coverage at different thresholds and show the precision-coverage plot for each analysis. We select fixed coverage based on system requirement and report relative precision numbers for each setting[7].

6.1 Effect of Question Category Classification

We evaluate three settings:

NoQC: All questions are treated as belonging to *specs* category and evaluated using ensemble matching model (refer to Section 3.4).

EnhancedQC: We use canned response for stock categories, ensemble matching model for three categories (*specs*, *price* and *compatibility* categories), and we provide no response for three remaining categories (*other*, *related_product* and *whats_in_box*).

StockQC: For this setting, we use the *EnhancedQC* classifier with a single difference that all non-stock categories are considered as *specs*. We use canned response for stock category questions and evaluate all non-stock categories using ensemble matching model. The motivation for this setting is that stock categories have limited surface forms and it is easier to detect them as compared to detecting non-stock categories which can have large number of surface forms.

Figure 3a shows precision-coverage curve for this analysis[8, 9]. At fixed coverage, *StockQC* setting achieves 14% improvement over *NoQC* setting. Using *EnhancedQC* setting, this improvement further increases to 19%. It can be concluded that significant improvement in precision is observed by having a simplified classifier that can detect stock categories and further precision improvements are observed by having an enhanced classifier (EnhancedQC setting).

6.2 Comparison of Matching Models

We study the performance of different matching models (as introduced in Section 3.4) against the *IDF-vector-average* baseline (refer to Figure 3b [10]. f_{annot} provides only one operating point whereas other models provide flexibility to choose operating point based on performance requirements. f_{deep}, by virtue of exploiting semantic signals in the training data, shows 85% better precision over the unsupervised *IDF-vector-average* baseline. Using cascade of f_{annot} and f_{deep} (refer to Section 3.4), this further improves to 93%.

6.3 Comparison of Data Sources

We analyze the contribution of three data sources, *product data*, *user reviews* and *CQnA*, on the system performance. We restrict this analysis to only specs questions and f_{deep} model. For *CQnA* evaluation, we obtain two similarity scores based on (1) user question and *CQnA* question similarity and (2) user question and *CQnA* answer similarity. We take maximum of these two similarity scores for selecting answer from *CQnA* data. For the evaluation with combined data setting, we use one answer from each source and a question is considered as positively answered if at least one answer is relevant.

As can be observed (refer to Figure 3c[11]), best results are obtained from the product data whereas performance on reviews and *CQnA* are comparable. By combining answers from these three sources, the performance improves drastically (11% improvement as compared to answer only from product data).

7 CONCLUSION

We presented a novel end-to-end framework for answering questions on e-commerce product pages. Based on this framework, we propose a question answering system, which uses deep learning-based ranking model and ontology-based matching to answer question from three sources—product data, reviews and CQnA. A CNN-based question intent classifier helps in further improving the precision of the system. Our proposed system, using question classifier and cascade of deep learning-based ranking and annotation-based matching, achieves 66% higher precision as compared to *IDF-vector-average* baseline.

[5]https://github.com/facebookresearch/fastText

[6]https://keras.io/backend/

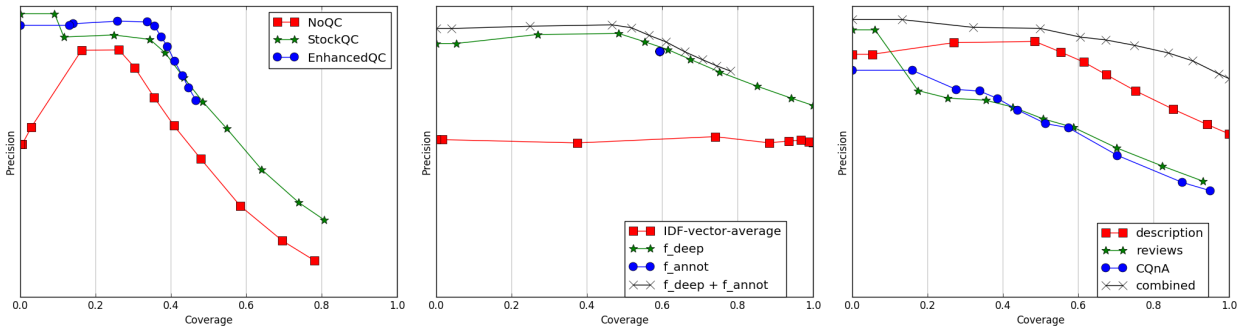
[7]Exact precision and coverage values are not disclosed due to confidentiality

[8]Coverage does not go till 100% due to f_{annot} filtering and filtering of question categories for *EnhancedQC* setting

[9]The *NoQC* setting has very low precision at low coverage, likely due to poor performance of deep learning model for questions belonging to "other" category

[10]f_{deep} has very low precision at low coverage due to errors introduce by incorrect parsing of data (e.g. "From the Manufacturer." answer for "who is the manufacturer?" question gets a score of 1 due to perfect word match after stopwords removal)

[11]The curves for reviews and *CQnA* data source do not go till coverage of 100% as reviews and *CQnA* are not available for newly introduced products



(a) (b) (c)

Figure 3: (a) Precision-coverage curve for different question classifier settings as evaluated on all test questions; (b) Precision-coverage curve for different matching models as evaluated on specs questions in test dataset; (c) Precision-coverage curve for different data sources as evaluated on specs questions in test dataset.

REFERENCES

[1] S. Arora, Y. Liang, and T. Ma. 2016. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. (2016).

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* (2016).

[3] Lei Cui, Furu Wei, Shaohan Huang, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A Customer Service Chatbot for E-commerce Websites, In Proceedings of ACL 2017, System Demonstrations. 97–102.

[4] Y. LeCun E. Sackinger J. Bromley, I. Guyon and R. Shah. 1994. Signature Verification using a Siamese Time Delay Neural Network. In *NIPS*.

[5] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[6] Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1408.5882* (2014).

[7] S. Lee and Y. Hu. 2015. Joint Embedding of Query and Ad by Leveraging Implicit Feedback. In *EMNLP*.

[8] X. Li. 2003. Syntactic Features in Question Answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

[9] B. Xiang M. Tan, C. Santos and B. Zhou. 2015. LSTM-based Deep Learning Models for Non-Factoid Answer Selection. *arXiv preprint arXiv:1511.04108* (2015).

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781* (2013).

[11] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv preprint arXiv:1611.09268* (2016).

[12] F. Peng, R. Weischedel, A. Licuanan, and J. Xu. 2005. Combining Deep Linguistics Analysis and Surface Pattern Learning: A Hybrid Approach to Chinese Definitional Question Answering. In *HLT/EMNLP*.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *CoRR* (2016).

[14] D. Ravichandran and E. Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. In *ACL*.

[15] D. Ravichandran, A. Ittycheriah, and S. Roukos. 2003. Automatic Derivation of Surface Text Patterns for a Maximum Entropy based Question Answering System. In *HLT-NAACL–Short Papers*.

[16] C. N. Santos, B. Xiang, and B. Zhou. 2015. Classifying Relations by Ranking with Convolutional Neural Networks. *arXiv preprint arXiv:1504.06580* (2015).

[17] R. Soricut and E. Brill. 2006. Automatic Question Answering using The Web: Beyond the Factoid. *Information Retrieval* (2006).

[18] E. M. Voorhees. 2001. The TREC Question Answering Track. *Natural Language Engineering* (2001).

[19] D. Wang and E. Nyberg. 2015. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. In *ACL-IJCNLP (Volume 2: Short Papers)*.

[20] M. Wang. 2006. A Survey of Answer Extraction Techniques in Factoid Question Answering. *Computational Linguistics* (2006).

[21] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. 2015. Towards Universal Paraphrastic Sentence Embeddings. *arXiv preprint arXiv:1511.08198* (2015).

[22] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *AAAI*.

Behavior Matching between Different Domains based on Canonical Correlation Analysis

Lyu Yuan
lyu.yuan@ist.osaka-u.ac.jp
Osaka University
Suita-shi, Osaka, Japan

Daichi Amagata
amagata.daichi@ist.osaka-u.ac.jp
Osaka University
Suita-shi, Osaka, Japan

Takuya Maekawa
maekawa@ist.osaka-u.ac.jp
Osaka University
Suita-shi, Osaka, Japan

Takahiro Hara
hara@ist.osaka-u.ac.jp
Osaka University
Suita-shi, Osaka, Japan

Niu Hao
ha-niu@kddi-research.jp
KDDI Research, Inc.
Chiyoda-ku, Tokyo, Japan

Kei Yonekawa
ke-yonekawa@kddi-research.jp
KDDI Research, Inc.
Chiyoda-ku, Tokyo, Japan

Mori Kurokawa
mo-kurokawa@kddi-research.jp
KDDI Research, Inc.
Chiyoda-ku, Tokyo, Japan

ABSTRACT

With the recent proliferation of e-commerce services, online shopping has become more and more popular among customers. Because it is necessary to recommend proper items to customers, to improve the accuracy of recommendation, high-performance recommender systems are required. However, current recommender systems are mainly based on information of their own domain, resulting in low accurate recommendation for customers with limited purchasing histories. The accuracy may suffer due to a lack of information. In order to use information from other domains, it is necessary to associate behaviors in different domains of the behaviorally related users. This paper presents a preliminary analysis of matching behaviors of the behaviorally related users in different domains. The result shows that we got a better prediction rate than linear regression.

CCS CONCEPTS

• **Information systems → Association rules**.

KEYWORDS

behavior matching, canonical correlation analysis, multi-domains

ACM Reference Format:
Lyu Yuan, Daichi Amagata, Takuya Maekawa, Takahiro Hara, Niu Hao, Kei Yonekawa, and Mori Kurokawa. 2019. Behavior Matching between Different Domains based on Canonical Correlation Analysis. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316595

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316595

1 INTRODUCTION

Online shopping has nowadays become increasingly popular. Due to the abundance of items, however, customers may have difficulty in making selections among them. To avoid such situations, online shopping systems employ item recommendation. Normally, recommender systems utilize content-based recommendations [6]. Based on item information of user histories, user preference profiles are generated. First, some item features are used to represent the items. Next, some user history information is used to generate user preference profiles. Finally, by comparing the user preference profiles with the item features, the system recommends items most related to the users [10].

On the other hand, many tasks use collaborative filtering because users often get the best recommendations from someone with similar tastes and interests to themselves. Based on this phenomenon, collaborative filtering matches people with similar interests and helps make recommendations [15]. To be more specific, collaborative filtering is a method of filtering for information or patterns utilizing techniques involving collaboration among multiple agents, viewpoints, data sources, etc [17]. Tasks employing collaborative filtering typically have large data sets. Collaborative filtering methods have been used in many situations. Examples include sensing and monitoring of data for mineral exploration; environmental sensing over large areas with multiple sensors; financial data, such as that used in financial service institutions integrating numerous financial sources; and electronic commerce and web applications where the focus is on user data [19].

The above recommender systems are usually based on information from their own domains, meaning that the data for analysis is limited. For instance, assume that a customer has purchased diapers in a shopping site. Accordingly, the customer would have an interest in baby products, thus the recommender system of the shopping site would put baby products into his/her recommendation list. However, with the exception of baby products, no items can be added to the list. The reason is that the recommender system has no notion of his/her preferences in other situations, resulting

in a lack of information. For instance, if the customer has accessed web pages with titles of chocolates oranges in other domains, then we can know that the customer has a tendency to have interest in chocolates and fruits. However, the previous domain, the shopping site, would have no knowledge about that, thus its recommendation system cannot put chocolates and fruits in the recommendation list. Using information from other domains may solve this problem. A user usually has multiple accounts in different domains. In addition to an account on the shopping site, for example, we may assume that the above-mentioned user has various accounts in other domains. It is useful for recommender systems if they can obtain information from such varied domains. However, the shopping site would typically have no knowledge of individual users' account IDs in other domains. As a result, though there may be abundant information on behaviorally related users in other domains, it is impossible for recommender systems to utilize such information. Therefore, matching the behaviors of the behaviorally related users in different domains would offer a solution for this problem.

In this paper, we consider the following situation. A user has accounts in two domains. One of the domains is an e-commerce site. In this domain, there are descriptions (mainly text) of products and purchase histories of users. The other domain is an Ad-Network service. The Ad-Network service puts advertisement on various webpages, and thus can record the URLs that users have accessed. Our task is to match the behaviors of the behaviorally related users in these two domains, by only using the above information.

The proposed method comprises the following steps. First, we create vectors for the accounts, based on the text information of the users. Second, since vectors from different domains are difficult to compare, we map these vectors into a shared space. After the vectors are mapped, we match the accounts based on similarity.

In our experiments, we evaluate canonical correlation analysis compared with linear regression. The results show that the former achieves better precision than the latter. Further experiments show that cosine similarity is more proper than Euclidean distance.

The rest of the paper is organized as follows. In Section 2, we introduce some related work. In Section 3, we describe our problem and situation. In Sections 4 and 5, we described our proposed method. In Section 6, we summarize our main experiment and results.

2 RELATED WORK

Since cross-domain matching uses information from other domains, it can be regarded as a problem of domain adaptation [4]. Domain adaptation is a problem utilizing machine learning and transfer learning. This problem occurs when we seek to learn from a source data distribution and a model with good performance on a different but related target data distribution. A representative task is picture recognition. In this task, the source distribution might be a model of person recognition, which must be adapted into a target distribution of animal recognition.

There are several types of domain adaptation. They differ in the information given in the target distribution. In unsupervised domain adaptation, the learning sample contains a set of labeled source examples, while a set of unlabeled source examples and an unlabeled set of target examples are given in learning and prediction [18]. In supervised domain adaptation, all the samples are labeled [8]. Since labeled examples are difficult to obtain, it is not realistic in most situations. To solve this problem, semi-supervised domain adaptation is utilized. In this situation, a small set of labeled examples in the target domain are used [7].

In this work, we use canonical correlation analysis to map vectors into a shared space, while the above works use Hierarchical Bayesian Model, and do not consider vector similarity.

3 PROBLEM STATEMENT

In this paper, we consider the following problem. There are two domains. One is an e-commerce site, referred to as Domain A in this paper. The other is an Ad-Network service, called as Domain B in this paper.

For Domain A, many items belonging to several categories are for sale on a shopping site. For every item, several kinds of information is given, such as price, text descriptions, category, selling duration, etc. Meanwhile, user information is also given, including the user's profile, such as gender, email address, username, etc. The user purchase histories are also available, such as purchased item, time, amount, price, etc. For Domain B, when opening a URL including an advertisement provided by this Ad-Network service, the advertisement will be displayed on its web page. When users access this advertisement, information on them is recorded, such as their IP address, access time, accessed URLs, etc.

We assume that users have accounts in both two domains. Each user has only one account in Domain A, but may have more than one account in Domain B. This is because the accounts in Domain B automatically update at a certain period of time. That is, the user automatically gets an updated account at the certain period of time. The problem in this paper is to match behaviors belonging to the behaviorally related users, by only using the information listed above. Additionally, to solve the problem, we only utilize the purchase histories in Domain A and the URL access log in Domain B. The structure of the domains is shown in Figure 1.

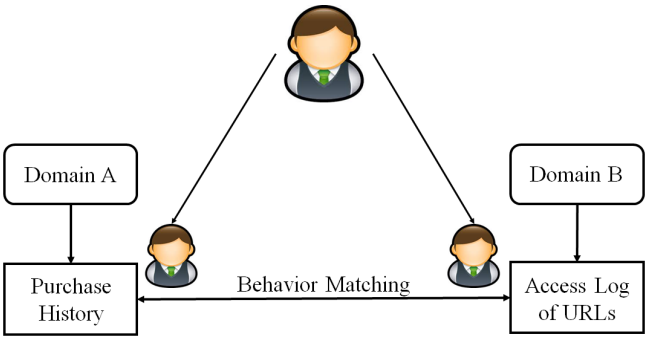


Figure 1: Domain structure

4 VECTOR GENERATION

For behavior matching, comparing the similarity of vectors of users between different domains may be the most practical method. The most meaningful information respecting the purchase histories and

URL access log is the articles involved. For example, the articles in the purchase histories may be item descriptions, while those in the URLs may be news articles. After we obtain articles, we use the obtained articles to generate user features.

4.1 Obtainment of articles

The URL access logs and purchase histories include articles, which are mainly obtained by crawling with special tags and in specific parts.

In Domain A, we have information on the purchase histories of users. We assume that items purchased by a user can represent the features of the user. Therefore, we use description articles of each item that the user has purchased, to obtain the users' features. For items in Domain A, the item title, promotion slogan and detailed description are combined as their articles.

In Domain B, we have the user access logs, in which the accessed URLs are recorded. For most URLs, web pages include articles, which could represent user preferences, and thus can be utilized to obtain their features. In HTML source files, since articles are usually included in tag *a* and/or tag *p* [16], we do crawling based on tag *a*, tag *p*, and the page title. In Web pages, tag *a* and tag *p* usually constitute the body, and thus include the main articles. The article of a Web page combines these three components.

4.2 Converting articles to vectors

After we have obtained articles from these two domains, we convert the articles into vectors. We utilize MeCab [1] for text segmentation, and create vectors using Doc2Vec [13]. The two procedures are shown below.

MeCab is an open-source text segmentation library for use with text in the Japanese language. It can analyze and segment a sentence into its parts of speech. Doc2Vec requires a list of words to make vectors, for which a long article is therefore unsuitable. Thus, it is necessary to divide articles into several words using MeCab.

The goal of Doc2Vec is to create a numerical representation of a document, regardless of its length. It learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Based on Word2Vec, it represents each document by a dense vector which is trained to predict words in the document [14].

4.3 Vectors of user accounts

After we have converted the articles into vectors, we create vectors for each user's account based on these article vectors. In Domain A, many users have purchased more than one item. In Domain B, all the users have accessed more than one URL. Thus, every account, both Domain A and Domain B, can have more than one article vector. Therefore, it is necessary to generate one account vector from several article vectors.

We assume that account r has a set of s articles, $\boldsymbol{A}_r = \{a_{r1}, a_{r2}, a_{r3}, ..., a_{rs}\}$, where a_{ri} is the i-th article of account r. Additionally, by converting the articles into vectors, the set of articles $\boldsymbol{A}_r$ has a corresponding set of converted vectors: $\boldsymbol{T}_r = \{\boldsymbol{t}_{r1}, \boldsymbol{t}_{r2}, \boldsymbol{t}_{r3}, ..., \boldsymbol{t}_{rs}\}$, where $\boldsymbol{t}_{ri}$ is the vector of the i-th article of account r. Therefore, $\boldsymbol{t}_r$, the vector of account r, is calculated as:

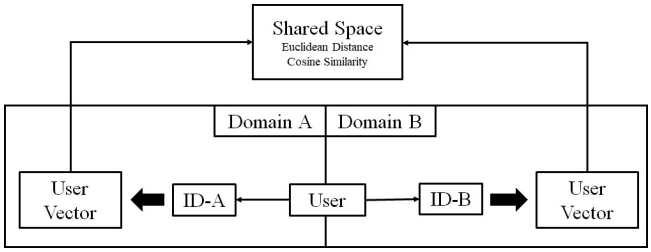


Figure 2: Matching System: the user has an account in both Domain A and Domain B. After obtaining vectors of the account, we map these into a shared space for matching.

$$\boldsymbol{t}_r = \frac{1}{s}\sum_{k=1}^{s} \boldsymbol{t}_{rk}. \tag{1}$$

That is, for a certain account, we use the average of all article vectors as its account vector.

5 BEHAVIOR MATCHING

After we have obtained account vectors of the two different domains, we match the accounts of the given user, based on these account vectors. A common solution is to match based on the similarity of vectors. However, since these vectors are generated from different domains, they would have different feature distributions. Therefore, it is impossible to employ vector similarity directly. Instead, it is necessary to map these vectors in a shared space. In this section, we first map account vectors into a shared space, and then calculate the similarity of the mapped vectors. Our matching system is shown in Figure 2.

A common solution is to calculate a conversion matrix for the two domains using linear regression. In this way, it is possible to use straight lines to fit the sample points [20]. We thus assume an n-dimensional vector X and an m-dimensional vector Y, where $\mathrm{X} = [x_1, x_2, x_3, ..., x_n]^\top$ and $\mathrm{Y} = [y_1, y_2, y_3, ..., y_m]^\top$. Both X and Y have several features, thus we want to analyze the relationship between X and Y. If $\mathrm{X} \in \mathbb{R}^n$ and $\mathrm{Y} \in \mathbb{R}^m$, we can establish the equation X = WY as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \tag{2}$$

where W is a conversion matrix for linear regression:

$$\mathrm{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix}. \tag{3}$$

Since $y_i = w_i^\top x$, it is necessary to train m times to get w_i of amount m, where y_i is the i-th feature of y and w_i is the i-th row in W. However, there is a problem in that each feature of Y is correlated to every feature of X, but has no correlation to other features of Y itself. To solve this problem, we utilize canonical

correlation analysis (CCA), which enables us to calculate multiple objective variables simultaneously.

5.1 Canonical correlation analysis

Canonical correlation analysis can be used as to determine the correlation between two pairs of vectors. If we have two vectors, $X = (x_1, x_2, x_3, ..., x_n)$ and $Y = (y_1, y_2, y_3, ..., y_m)$, canonical correlation analysis will use linear calculations to determine the combinations of X and Y which have maximum correlation with each other, and give a solution for the maximum correlation [9].

We here assume that X and Y are account vectors in different domains of the same user, and canonical correlation analysis then seeks vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ for the following situation [11]:

$$\max \quad \rho = \mathrm{corr}(\boldsymbol{a}^\top X, \boldsymbol{b}^\top Y). \tag{4}$$

By maximizing the correlation between $\boldsymbol{a}^\top X$ and $\boldsymbol{b}^\top Y$, we can convert the paired vectors X and Y into similar vectors X' and Y', as a result of mapping them into a shared space, where $X' = \boldsymbol{a}^\top X$ and $Y' = \boldsymbol{b}^\top Y$ [5]. In this way, we can calculate the similarity of vectors in the two different domains.

In canonical correlation analysis, the converted vectors X' and Y' usually have the same number of dimensions, which can be selected according to our needs. The greater the number of dimensions, the more features the vectors will have. However, greater numbers of dimensions may require longer time for training models [3].

5.2 Similarity calculation

After obtaining converted vectors, we can match vectors based on similarity. Typically, when calculating the similarity of vectors, Euclidean distance and cosine similarity are used. In mathematics, Euclidean distance is the ordinary straight-line distance between two points in Euclidean space. The Euclidean distance between X and Y is calculated as:

$$\text{distance} = d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{5}$$

Cosine similarity is a measure of the similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. For vectors X and Y, the cosine similarity is calculated as:

$$\text{similarity} = \cos(\theta) = \frac{X \cdot Y}{||X||||Y||}, \tag{6}$$

where ||X|| and ||Y|| are the magnitude of vectors X and Y.

6 EXPERIMENT AND RESULTS

Our experiment is based on real-world databases of two different domains. The user has accounts in both Domain A and Domain B. In Domain A, the user has only one account, and in Domain B, at least one account.

6.1 Data set

We utilized real data[1] collected from May 11$^{\text{th}}$ 2017 to March 31$^{\text{th}}$ 2018. The dataset includes 82 thousand users. Therefore, in Domain A, there are 82 thousand accounts, for which there is at least one purchase record. Meanwhile, in Domain B, since one user could have more than one account, there are 92 thousand accounts, for each of which there are several URL access records.

For each of the URLs accessed in Domain B, we calculate the access time. Since there are more than 11 million accessed URLs, it is almost impossible to do crawling for all the articles. Therefore, we only crawled articles of URLs with more than 1,000 accesses.

After we obtained all the needed articles from the two domains, we created article vectors using MeCab and Doc2Vec, as outlined above. Vectors were created in 256-dimension.

[1]Due to the policy of the data providers, we cannot reveal the data sources.

6.2 Evaluation methodology

For the evaluation metric of our experiment, we used Precision@R, which involves matching performance based on using the precision with which the true target instance is included in a set of R proportion of candidate instances, $S(R)$, found by each method. More formally, the precision is given by:

$$\text{Precision@}R = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \delta(t_i \in S_i(R)), \tag{7}$$

where N_{te} is the number of test instances in the target domain, t_i is the i-th true target instance, $S_i(R)$ is the R candidate instances of the i-th source instance and $\delta(\cdot)$ is the binary function that returns 1 if the argument is true, and 0 otherwise. Here, R means ratio in the dataset, but not an integer number.

In our experiments, we used 70% of the data to train our model and the remainder (30%) for testing. For each condition, we performed the experiment five times with different allocations of training and testing data. After all the experiments under the same condition were completed, we averaged the results of each experiment as the final result.

6.3 Similarity calculation

Since we have two options (cosine similarity and Euclidean distance) to calculate similarity, we need to determine which is suitable for our experiment. We tested under many conditions, and determined that cosine similarity is more suitable. An example is shown in Figure 3, where the dimension of mapped vectors is nine.

6.4 Dimension of mapped vectors

When training the model using canonical correlation analysis, we must set the parameter of the dimension of mapped vectors. A larger dimension means more user features, but also has negative effects, such as more calculation consumption and feature noise.

We tested using parameters of dimension of mapped vectors from 1 to 15. All the results of the canonical correlation analysis are shown in Tables 1 and 2, using both Euclidean distance and cosine similarity. The results in Tables 1 and 2 on dimension one are the same, since the calculation of cosine similarity and Euclidean distance is the same on dimension one. Bold values are the best performance given a specific dimension of mapped vectors. When the dimension of mapped vectors is near nine, the performance has a peak value. Since we selected cosine similarity, we determined the best dimension using cosine similarity.

Table 1: Result on Euclidean distance of number of dimension from 1 to 15

R	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.05	0.075	0.088	0.094	0.101	0.104	0.108	0.111	0.112	0.113	0.113	0.113	0.113	0.113	0.113	**0.114**
0.1	0.142	0.165	0.177	0.187	0.193	0.198	0.201	0.201	0.201	**0.202**	0.202	0.201	0.201	0.201	0.200
0.15	0.205	0.236	0.250	0.261	0.267	0.272	0.274	0.274	0.274	**0.274**	0.273	0.272	0.271	0.271	0.270
0.2	0.263	0.300	0.314	0.325	0.331	0.334	**0.336**	0.335	0.334	0.333	0.332	0.332	0.331	0.330	0.329
0.25	0.316	0.359	0.373	0.383	0.388	**0.391**	0.391	0.391	0.389	0.388	0.387	0.386	0.385	0.383	0.382
0.3	0.369	0.416	0.429	0.438	0.443	**0.444**	0.443	0.442	0.442	0.440	0.439	0.438	0.436	0.435	0.433
0.35	0.421	0.470	0.483	0.490	0.494	**0.494**	0.494	0.492	0.491	0.490	0.489	0.488	0.487	0.485	0.484
0.4	0.473	0.522	0.534	0.540	0.543	**0.543**	0.542	0.541	0.540	0.538	0.536	0.535	0.535	0.533	0.532
0.45	0.523	0.572	0.584	0.589	0.591	**0.591**	0.590	0.589	0.587	0.585	0.584	0.583	0.582	0.580	0.578
0.5	0.572	0.622	0.633	0.636	**0.637**	0.637	0.636	0.634	0.633	0.631	0.630	0.629	0.628	0.626	0.625

Table 2: Result on cosine similarity of number of dimension from 1 to 15

R	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.05	0.075	0.102	0.120	0.130	0.135	0.143	0.145	0.148	0.148	0.149	**0.150**	0.150	0.149	0.149	0.150
0.1	0.142	0.207	0.229	0.240	0.247	0.255	0.258	0.259	0.261	**0.261**	0.261	0.261	0.261	0.261	0.261
0.15	0.205	0.290	0.314	0.325	0.332	0.339	0.343	0.345	**0.347**	0.346	0.346	0.346	0.346	0.345	0.345
0.2	0.263	0.356	0.381	0.392	0.399	0.406	0.410	0.412	**0.413**	0.413	0.412	0.412	0.412	0.412	0.410
0.25	0.316	0.415	0.440	0.452	0.459	0.466	0.470	0.472	**0.472**	0.472	0.472	0.472	0.471	0.471	0.471
0.3	0.369	0.471	0.496	0.509	0.516	0.522	0.526	0.527	**0.528**	0.528	0.527	0.526	0.527	0.526	0.525
0.35	0.421	0.524	0.550	0.562	0.568	0.574	0.577	0.578	**0.580**	0.579	0.579	0.579	0.578	0.577	0.576
0.4	0.473	0.575	0.600	0.610	0.617	0.622	0.625	0.626	0.626	0.626	**0.626**	0.626	0.625	0.624	0.624
0.45	0.523	0.624	0.646	0.657	0.661	0.667	0.670	0.671	0.671	0.671	**0.671**	0.670	0.670	0.670	0.669
0.5	0.572	0.669	0.690	0.700	0.704	0.710	0.712	0.714	**0.714**	0.713	0.712	0.712	0.712	0.712	0.711

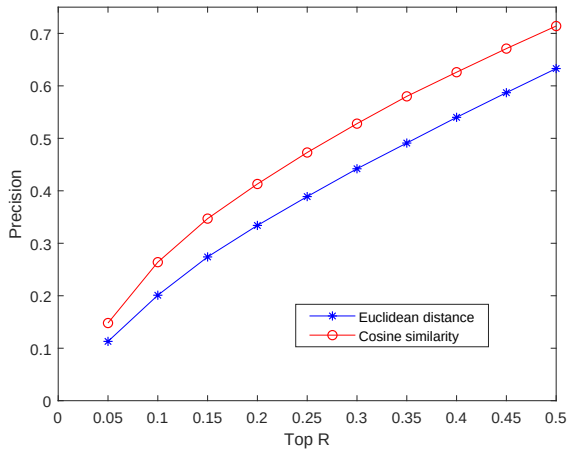


Figure 3: Similarity calculation: for each top R proportion of candidates, CCA's precision with Euclidean distance and cosine similarity.

6.5 Comparison with linear regression

Linear regression is a baseline method in our problem, and is also employed, for comparison with the proposed. Our method used nine dimensions for the account vectors. Both two experiments used cosine similarity. The results are shown in Figure 4. It is obvious that our method outperforms the baseline method.

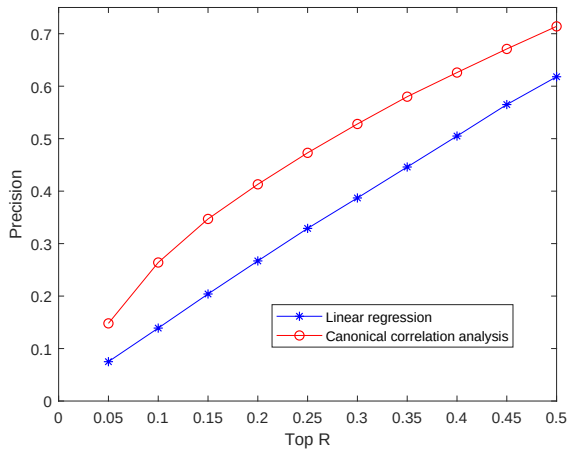


Figure 4: Comparison with linear regression: for each top R proportion of candidates, CCA's precision compared with linear regression.

7 CONCLUSION

In this study, we developed a system to match behaviors in different domains, for the same users. In order to use vector similarity to match the behaviors, we generate vectors based on articles related to users, by using MeCab and Doc2Vec. When matching behaviors, we have found that canonical correlation analysis and cosine similarity show better performance results than linear regression. Actually, our result is a preliminary and the precision is not high enough for practical applications. Therefore, as for future work, we plan to utilize some non-linear methods, such as deep canonical correlation analysis [2] and kernel canonical correlation analysis [12].

ACKNOWLEDGMENT

This research is partially supported by JST CREST Grant Number J181401085.

REFERENCES

[1] 2013. Mecab. http://taku910.github.io/mecab/.
[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proc. 2013 International Conference on Machine Learning*. 1247–1255.
[3] Francis R Bach and Michael I Jordan. Technical report, University of California, Berkeley, 2005. A probabilistic interpretation of canonical correlation analysis. (Technical report, University of California, Berkeley, 2005).
[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. 2010. A theory of learning from different domains. *Machine Learning* 79, 151–175.
[5] Patricia Cohen, Stephen G West, and Leona S Aiken. 2014. *Applied multiple regression/correlation analysis for the behavioral sciences*.
[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proc. 10th ACM Conference on Recommender Systems*. 191–198.
[7] Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proc. the 2010 Workshop on Domain Adaptation for Natural Language Processing*. 53–59.
[8] Daniel Garcia-Romero and Alan McCree. 2014. Supervised domain adaptation for i-vector based speaker recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4047–4051.
[9] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16, 12 (2004), 2639–2664.
[10] Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A Neural Probabilistic Model for Context Based Citation Recommendation. In *Proc. 29th AAAI Conference on Artificial Intelligence*. 2404–2410.
[11] Thomas R Knapp. 1978. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin* 85, 2 (1978), 410.
[12] Pei Ling Lai and Colin Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* 10, 05 (2000), 365–377.
[13] Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368* (2016).
[14] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. 2014 International Conference on Machine Learning*. 1188–1196.
[15] Dongsheng Li, Chao Chen, Qin Lv, Hansu Gu, Tun Lu, Li Shang, Ning Gu, and Stephen M. Chu. 2018. AdaError: An Adaptive Learning Rate Method for Matrix Approximation-based Collaborative Filtering. In *Proc. 2018 World Wide Web Conference*. 741–751.
[16] Gengxin Miao, Junichi Tatemura, Wang-Pin Hsiung, Arsany Sawires, and Louise E Moser. 2009. Extracting data records from the web using tag path clustering. In *Proc. the 18th International Conference on World Wide Web*. 981–990.
[17] Shakiba Rahimiaghdam, Pinar Karagoz, and Alev Mutlu. 2016. Personalized Time-aware Outdoor Activity Recommendation System. In *Proc. 31st Annual ACM Symposium on Applied Computing*. 1121–1126.
[18] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric Tri-training for Unsupervised Domain Adaptation. In *Proc. 2017 International Conference on Machine Learning*.
[19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based Collaborative Filtering Recommendation Algorithms. In *Proc. 10th International Conference on World Wide Web*. 285–295.
[20] George AF Seber and Alan J Lee. 2012. *Linear regression analysis*. Vol. 329.

Neural Product Retrieval at Walmart.com

Alessandro Magnani
Walmart Labs
Sunnyvale, United States
amagnani@walmartlabs.com

Feng Liu
Walmart Labs
Sunnyvale, United States
feng.liu@walmartlabs.com

Min Xie
Walmart Labs
Sunnyvale, United States
mxie@walmartlabs.com

Somnath Banerjee
Walmart Labs
Sunnyvale, United States
sbanerjee1@walmartlabs.com

ABSTRACT

For an E-commerce website like Walmart.com, search is one of the most critical channel for engaging customer. Most existing works on search are composed of two steps, a *retrieval* step which obtains the candidate set of matching items, and a *re-rank* step which focuses on fine-tuning the ranking of candidate items. Inspired by latest works in the domain of neural information retrieval (NIR), we discuss in this work our exploration of various product retrieval models which are trained on search log data. We discuss a set of lessons learned in our empirical result section, and these results can be applied to any product search engine which aims at learning a good product retrieval model based on search log data.

KEYWORDS

Product Search; Neural Information Retrieval

ACM Reference Format:
Alessandro Magnani, Feng Liu, Min Xie, and Somnath Banerjee. 2019. Neural Product Retrieval at Walmart.com. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316603

1 INTRODUCTION

Search is one of the most important channels for customers to discover products on an E-commerce website such as Walmart.com. Given our huge catalog which contains more than 100 millions products, how to help user surface the right products based on their query becomes a very challenging problem.

Existing literature on information retrieval focuses mostly on web search [18]. While product search shares many common challenges as with web search, there are many aspects in product search which are unique. For example, web search focuses on retrieving documents that satisfy a userâĂŹs *information need*, e.g., when searching for *when is the next olympic*, any web page from trusted source which describes this information will satisfy the user's need; whereas in product search, it serves users' *shopping need*, among many assorted products which match users' query, user will likely only buy one, thus the product search engine has to identify which one is the best.

Similar to existing general search engine, product search usually involves two steps, the first step is to *retrieve* all relevant products from the catalog which form the recall set; then these candidate products will go through a *re-rank* step to identify which products are the best to return to the customer. Most of the recent learning to rank efforts focus on the re-ranking step, and they leverage extensive features available and can potentially involve very complicated ranking model [17]. In the retrieval step, query is potentially matched against every document in the collection thus latency of the retrieval model is a big challenge, and we usually have to trade-off between model complexity and scalability of the retrieval model. In this work, we focus on the retrieval step.

For a classical search engine, retrieval is mostly based on the raw text and a heuristic score function such as Okapi BM25 [18]. Inverted index as implemented in Apache Lucene [1] can be leveraged to make retrieval fast. Matching product and query in this solution is based on exact matching of tokens, which may not be suitable for product search because of the vocabulary mis-match between query and product as found by many recent works [11, 27]. E.g., users might search for a u-disk, whereas the matching items in the catalog are named after flash drive. Many existing works are trying to solve this problem by incorporating knowledge graph [31] or having a dedicated query understanding component [9]. However, these approaches need huge amount of domain expertise, and the cost of maintaining these components is very high as the product vocabulary and the catalog of the E-commerce websites change frequently.

In this work, inspired by recent progress in *neural information retrieval* (NIR) [20], we consider a few end-to-end neural product retrieval solutions which can be trained directly on abundant search query log which is available to any product search engine.

The first model we consider builds its foundation directly on the raw text of the underlying query and product. This model captures directly the text level similarity between query and product, then applies neural transformation on the shared text representation to predict the relevance between query and product. While non-linear transformation can capture the more complicated correlation between text in query and product, still this approach suffers from the need of exact matching of tokens in the text.

Our second model leverages a semantic embedding layer to first transform texts of query and product to their corresponding high dimensional embedding representation, then we apply a neural

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316603

transformation on the concatenated query and product embedding vector. This model is inspired by recent advancement in word embedding [19], and has the most powerful modeling capacity. However, when applying to the search retrieval problem, applying the neural transformation to query and each product in a huge catalog can be prohibitive.

Finally, we propose a neural retrieval model which applies neural transformation to obtain vector representation on product and query separately, and then apply a simple aggregation function to combine the two vectors to predict the final score. Unlike the second approach, scoring query and product representation in the neural retrieval model can be as simple as vector multiplication. Thus it can be handled efficiently in the product search engine based on Apache Lucene [1], or use fast approximate nearest neighbor search library such as Faiss [15].

We also discuss in our work how we leverage a character level tokenization method inspired by [11, 27], how different rank loss functions such as point-wise loss and pair-wise loss can be implemented in our framework, and how we can incorporate additional product level attribute information such as brand, price, click-through rate which are critical to the search ranking problem into our model. We share in our experiment results section lessons we learned by applying our proposed solutions on a big search log dataset from Walmart.com. We believe these results can be applied to any product search engine which aims at learning a good product retrieval model based on search log data.

The following of the paper is organized as follows: we first discuss a few related works in Section 2, then we present the proposed product retrieval models in Section 3. In Section 4 we present our empirical evaluation results of our proposed solutions. And finally we conclude and discuss future works in Section 5.

2 RELATED WORK

In [26], the authors discuss that product search in E-commerce requires specialized solutions and challenges in ranking products. In [8], the authors propose a keyword based retrieval solution for product search, the idea is to generate keywords related to each product based on query logs and reviews, and then index these keywords based on inverted index. Unlike our end-to-end solution, generating keywords for products can be tricky as it involves tuning of various thresholds and domain expertise. Brenner et al. studies a end-to-end product retrieval problem on Jet.com in [4]. In this work the authors explore a similar setting as our neural retrieval model as discussed in Section 3, however, unlike our work which leverages the abundantly available raw query product engagement data, the authors explore labels which are obtained from query refinement data which are inferred from online user sessions. Though potentially the quality of this data is high, the challenge of this approach lies in the fact that the inferred query refinement data is very sparse. Some recent efforts in product search explore some interesting but orthogonal directions such as multiple objective [29] and product substitutability [10].

Neural information retrieval (NIR) has been a popular topic in the search community recently. It involves a set of sub-topics such as unsupervised learning of text embedding such as word2vec [19], deep autoencoder-based model such as semantic hashing [24], deep

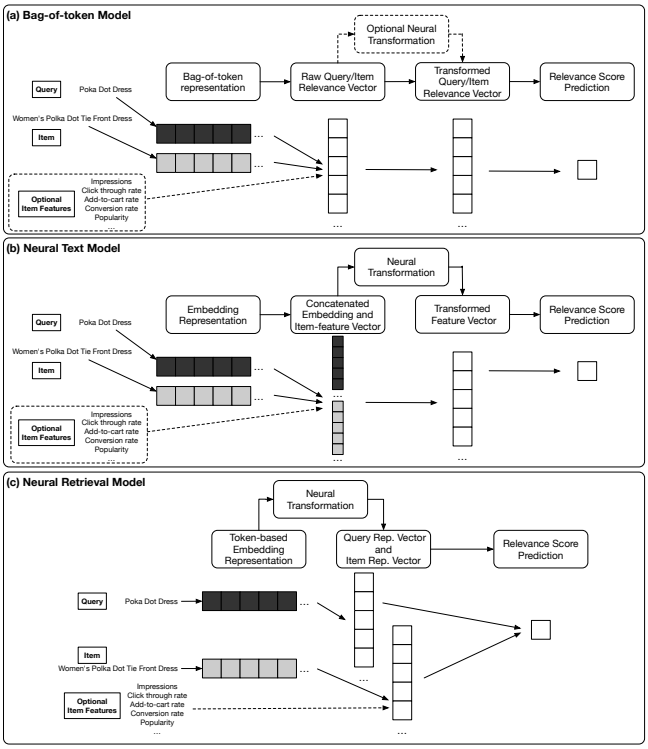


Figure 1: Models for neural product retrieval

siamese model based on query log such as CLSM [27] and DSSM [11], and query document interaction-based model such as kernel pooling [30]. A good summary of the field can be obtained in [20]. Our neural retrieval model in Section 3 is closely related to CLSM and DSSM. Similar to [4], we are focusing more on the problem of product retrieval, and exploring solution which can strike a good balance between retrieval performance and model complexity.

3 NEURAL PRODUCT RETRIEVAL

To solve the problem of product retrieval, the key is to assess the relevance between a query and a product. In this section we propose three different models: A simple bag-of-token model based directly on tokens in query and product, a neural text model which can incorporate semantic similarity between query and product, and lastly a customized neural text model which is more optimized for retrieval. We also discuss how we tokenize text information of query and product for processing, how we can incorporate additional critical product information such as product price, brand and click-through rate into the model, and the different loss functions we have tested for the model.

3.1 Token Representation for Text

One simple way to represent text is to consider it as a bag of words which are contained in the text. The problem with this approach is that usually the vocabulary size is large, and this approach is difficult to handle mis-typed words in user typed queries.

In this work, we consider a token representation approach at sub-word level based on letter-trigram (LTG), which is inspired

by [11, 27]. In this approach, for an example query "red dress", we first derive the letter-trigrams from each word, e.g., "red" will be broken down to three letter-trigrams, "#re", "red", "ed#", where # is the special word boundary character added to each word; and then the overall text will be represented as a bag of letter-trigrams. Our approach is also in part inspired by recent effort on subword embedding which demonstrated good performance in a set of natual language processing tasks [3].

3.2 Bag-of-Token Model

Let q be a query, and p be the text information about a product, based on the discussion from previous section, we can represent q and p as bags of tokens, where tokens are formed of letter-trigrams. Then we can apply one-hot encoding to each token and obtain a dense count vector for q and p as $q = [q_1, \ldots, q_n], p = [p_1, \ldots, p_n]$, where each q_i/p_i is the count of the corresponding token appears in q/p, and n is the size of the vocabulary.

A basic bag-of-token based scoring function can feed the query product pair through a simple linear model as below, where w and b are weight parameter and bias term of the model, and $\circ$ is element-wise multiplication.

$$\mathcal{F}_{BOT}^{linear}(q, p) = w \cdot (p \circ q) + b \tag{1}$$

We call this model BOT, short for Bag-of-tokens model. Essentially, this basic model examines the shared tokens between query and product, then it makes prediction based on this shared information using a linear model. This approach can be thought of as a simplified version of the vector space model for information retrieval [25].

The basic model above does not have the capacity to model the complex interaction between tokens and their correlation with the label. Thus we extend the basic approach by replacing the linear scoring component with a neural network model which is shown in Figure 2. In this model, we feed the aggregated count vector between query and product through N hidden layers of a neural network in the forward pass, Specifically, each hidden layer computes:

$$a^{(l+1)} = \phi(W^{(l)}a^{(l)} + b^{(l)}) \tag{2}$$

where l is the layer number and ϕ is the ReLU activation function [21]. $a^{(l)}$, $b^{(l)}$, and $W^{(l)}$ are the activations, bias, and model weights at l-th layer. We also added batch normalization [12] after each dense layer, and a dropout layer [28] at the end. This approach is inspired by Deep Averaging Network [13], and we call this model BOT-N (Bag-of-tokens with Neural Transformation).

We note that though BOT-N can model more complex interactions between tokens in query and product, but text information on product alone, e.g., title, usually are not enough for distinguishing good products from bad products. Two products can have the exact same title like polka dot dress but huge gaps in terms of sales. Some structured attributes of products usually are very important for the search task, e.g., brand of the product, price of the product, click-through rate of the product, add-to-cart rate of the product, and number of orders of the product. Thus we consider extend above models with these additional product level features by appending the transformed product features as also the input to the model, we call the extended model BOT-N$^+$. A list of key product features

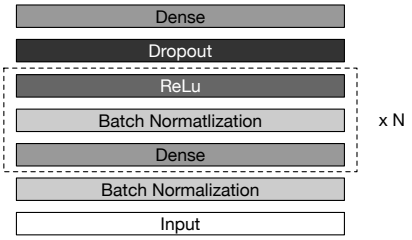


Figure 2: NN transformation

Feature Name	Feature Description
Brand	Brand for the product
PT	Product type for the product
Dept	Department of the product
Price	Price of the product
Popularity	Popularity of the product in terms of orders
CLK	Number of clicks for the prodoct
CTR	Click through rate for the product
ATC	Number of add to cart for the product
ATCR	Add to cart rate for the product
ORD	Number of orders for the product
ORDR	Order rate for the product

Table 1: Key product features used as model input.

used in this work is listed in Table (1), and all feature values in this work were collected in a 6 month window in 2018.

The architecture of the BOT model is shown in Figure 1 (a).

3.3 Neural Text Model

The bag-of-token models score query product pairs by directly considering tokens involved in the underlying query and product. The problem of this approach is that it has limited capability of modeling semantic similarity relationships between tokens, e.g., synonyms such as nightgown versus girls' pajama is difficult to be handled.

Inspired by recent advancement in text embedding [19], we consider a model architecture which first transform the input text to a embedding vector using either a simple convolutional neural network (CNN) similar to [16] or a simple embedding layer learned from scratch. For the simple embedding layer based approach, we use average pooling to obtain the embedding representation for the entire text piece, similar to deep averaging network [13] as used in universal sentence encoder [5]. In our experiment, we found that for our dataset mostly the simply embedding layer based approach will win against the more complicated CNN-based approach, and the CNN-based approach will take way longer time for training, thus we will skip the discussion of this approach in this work.

For a query product pair, we obtain the two corresponding embedding vectors, and then we concatenate the two embeddings and apply a neural transformation of the concatenated embedding vector through a few fully connected layers as shown in Figure 2.

We call this model Neural Text Model or NTM. Similar to the bag-of-token model, we can incorporate product features in the

the model as well for predicting query product relevance. The way we incorporate product feature is very similar to the bag-of-token model, we directly append the product features as input to the embeddings of query and product text, and apply neural transformation layer on top of all embedding the product features.

3.4 Neural Retrieval Models

As discussed in the introduction, BOT-N and NTM models all involve complex neural transformation based on the query and product information. This will introduce significant overhead when implementing the functionality in practical product retrieval application. This is because even though product features and embedding can be pre-computed, query tokens and embeddings have to be generated on the fly, and thus we have to score potentially all products againt query embedding through the more complicated transformation to find which products might have a better matching score w.r.t. the query. This is why even though more complicated learning to rank methods such as Kernel Pooling [30] has good performance on ranking, they can only be applied in the re-ranking phase when doing product search as discussed in [4].

Based on this intuition, we consider a modified neural text model named Neural Retrieval Model (NRM), in which we generate embedding representation of product and product separately, and feed both embeddings into a simple differentiable transformation function such as *cosine* or *dot-product*. The architecture of NRM is shown in Figure 1 (c). And similar to BOT and NTM, we can easily add product features as input when generating product embedding representation, we call this extended model NRM$^+$.

3.5 Model Training

We consider training our model directly on the abundant query log data from product search engine. Similar to many existing works in *learning-to-rank* [17], we consider point-wise approach and pair-wise approach for fitting our product retrieval model.

For the point-wise approach, given a pair of query q and product p, we can apply any of the models proposed in this section to obtain a score $\hat{y}_{q,p} = \mathcal{F}(q,p)$. Then in the loss function, we fit $\hat{y}_{q,p}$ through sigmoid cross entropy and against a label y which is the click-through rate of query q and product p.

$$L(\hat{Y}, Y) = -\sum_{q,p,y} y \times \log(\sigma(\hat{y}_{q,p})) + (1-y) \times \log(1-\sigma(\hat{y}_{q,p})) \quad (3)$$

For the pair-wise approach, we adopt the logistic pair loss defined below, where $\mathbb{I}$ is the indicator function.

$$L(\hat{Y}, Y) = \sum_{q,p_i,p_j,y} \mathbb{I}(\hat{y}_{q,p_i} > \hat{y}_{q,p_j}) \log(1 + exp(\hat{y}_{q,p_i} - \hat{y}_{q,p_j})) \quad (4)$$

To determine the training hyper parameters and to avoid overfitting, we divide the data into two sets that do not overlap, called training and validation datasets, respectively. Details of the data preparation will be discuss in Section 4.

We use Adagard as optimizer for training our model, and we set learning rate to be either 0.128 or 0.02 depending on setting. In our implementation, models are trained using a small GPU cluster with Nvidia Tesla V100.

3.6 Handling Position Bias

One challenge of leveraging implicit query log data is the position bias as discuss in recent literature [6]. Intuitively, products which are ranked higher by the existing algorithm will accumulate more clicks because they are more likely to be examined by customers; whereas products shown lower in the search result page will likely get less attention. In this work, we consider a simple click model which estimates the times that product is observed by a user only as a function of impressions and position. In our click model, the number of *examines* of a query q and a product p is estimated using the following formula:

$$\text{examine}_{q,p} = \sum_i \eta_i \times \text{impression}_i \quad (5)$$

In this equation, we sum over all positions i where impression_i is the number of impression at position i for query q and product p. We weight the number of impressions in the log data by a global position bias term η_i which is fitted empirically. Intuitively, η_i will weight the impression count lower if the position i is high, and it will weight the impression higher if the position i is low.

We then compute the click-through rate as the ratio of clicks and examines and order rate as the ratio of orders and examines. We can also use the examine number for each query product training instance to weight either the point-wise loss or pair-wise loss. This weighting scheme will capture both position bias, and also it will capture the confidence of each training instance, as training instance with more impressions/examines we are more confident on the data, and it will have a higher weight value.

4 EMPIRICAL EVALUATION

The implementation of our product search models is based TF-Ranking [22], which is a recently proposed open source library for solving large-scale ranking problems in Tensorflow [2]. In the following of this section, we first present the dataset used in our work, then we discuss metrics we used in our evaluation, and finally we present our findings by comparing the different product search models presented in this work.

4.1 Dataset

We collected a large query log dataset on shoes segment during a six month window from May 2018 to October 2018 on Walmart.com. Historical data of the extra features such as clicks and orders are collected from the query log six months before May 2018. The dataset is composed of more than 100 million query and product pairs, of which there are more than 1 million unique queries and more than 1 million unique item titles. We randomly split the data into training and validation dataset with the ratio 92.5%:7.5%, and we make sure there are no shared query and product pairs in the training and validation dataset.

4.2 Metrics

The first metric we use to evaluate our model is the standard ranking metric NDCG (Normalized Discounted Cumulative Gain) [14]. We

consider NDCG at different positions 3, 5, and 10.

$$\text{DCG} = \sum_{j=1}^{n} \frac{2^{y_j} - 1}{\log_2(1 + r_j)}, NDCG = \frac{1}{N} \sum_{i=1}^{N} \frac{DCG}{DCG_{\pi^*}} \quad (6)$$

Where r_j and y_j is the rank and the score respectively of item j. As score we use the ratio of orders and examines, where examines are defined as in Section 3.6. Note that though we are optimizing for order rate, we did not use order rate as label for training because order data is sparser compared with click data.

For product search we usually want to rank those products which have higher order rate at top, thus this inspired us to consider another metric Pair Accuracy which measures in the validation set that for a query q, of all product pairs p_i and p_j where p_i has a higher order rate compared with p_j, how many of them the model will predict correctly.

We also track some other metrics such as ARP (Average Relevance Position), the result is very similar to NDCG and Pair Accuracy. Thus we skip the discussion of those numbers in this paper.

4.3 Bag-of-Token Model

As expected the basic BOT model which is based solely on the shared tokens between query and product, does poorly on all metrics, this is understandable since the modeling power of this baseline without any transformation is very limited. Thus we use the metric values from BOT as our baseline, and we show only the improvement over BOT for other models on all metrics.

We show our experiment results of the BOT-N models in Table 2. As can be seen from this table, For BOT-N we added in our experiment three dense layers which serve as the neural transformation on the shared token information, and these added transformation layers can improve significantly on the performance over the basic BOT. Adding additional product level features in BOT-N$^+$ will help lift all metrics compared with BOT-N, which is again expected, as based on our experience, the added features are usually highly correlated with the metric we are evaluating.

4.4 Bag-of-Token Model vs. NTM Model

In Table 2, we also show the comparison of the more advanced neural text models with our bag-of-token models. From the table, it is very clear that the added token embedding layer boosts the metrics over the BOT significantly. This matches our intuition as these NTM models has more modeling capacity based on their learning of token embedding representation that can capture semantic similarity relationship between them. Also we can observe that similar to BOT-N vs. BOT-N$^+$, adding additional product features also helps on all the metrics we are tracking.

4.5 Pair Loss vs. Point Loss

We compare the pair-wise loss function and point-wise loss function in Table 3. And as can be easily observed from this table, the pair-wise loss function introduces a better performance compared with point-wise loss function. This should be intuitive as all metrics we are tracking are all based on ranking, it matches the direction of the pair-wise loss function better.

4.6 NTM Model vs. NRM Model

To compare pair loss with point loss, we show in Table 3 the performance comparison results for NTM and NRM. The difference between these two models essentially lie in the fact that whether we perform a complex transformation layer once query and product embedding representation have been obtained. Intuitively, NTM should perform better as the additionally introduced transformation layers can help matching query to its relevant products. However, as discussed in the previous sections, the important benefit of NRM lies in its efficiency gain in practical implementation. As query embedding and product embedding in NRM can be matched with simple operations, we can leverage fast high dimensional approximate nearest neighbor retrieval libraries such as Faiss [15]. As shown in Table 3, as expected, the metric performance with NRM is worse than the NTM models, however, it is not far behind, and still comparable to many models we considered in the experiment. We think NRM with pair loss strikes a good balance between modeling power and empirical performance.

4.7 Head Query vs. Tail Query

We also consider weighted NDCG and weighted pair accuracy metric which are weighted based on query traffic, thus naturally these weighted metrics are more leaning towards head queries, whereas the non-weighted metrics are more leaning towards tail queries. In Table 4, we show a comparison of weighted metrics vs. non-weighted metrics for NTM and NTM/Pair Loss. As can be seen from the table, it is very consistent that most of the case our models have a better performance on weighted metrics compared with the corresponding non-weighted metrics. This can be simply explained by the fact that in the training data, we weight training examples based on impression, thus naturally the model is putting heavier emphasis on the training instances with more query traffic.

5 CONCLUSION AND FUTURE WORK

Inspired by latest works in the domain of neural information retrieval (NIR), we discuss in this work our exploration of various product retrieval models which are trained on search log data. We present a set of lessons learned based on our exploration that can be easily applied to any product search engine.

In the future, we plan to improve our retrieval models by exploring multiple directions. Though letter trigram based tokenization has the benefit of having a small vocabulary and being able to naturally handle mis-typed user queries better, it cannot leverage word-level embedding pre-trained on large text datasets. We are exploring to incorporate the latest text embedding works such as ELMo [23] and BERT [7] into our models. Also while experimenting with our neural retrieval model, we noticed that the underlying neural network structure has heavy impact on the performance of the model, we plan to investigate more on this topic in the future. Finally we also plan to explore the use of other click models to better capture user interaction with the site.

REFERENCES

[1] 2019. Apache Lucene. http://lucene.apache.org.
[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray,

Model	NDCG@3	NDCG@5	NDCG@10	Pair Acc.
BOT-N	19.84%	19.00%	15.33%	3.83%
BOT-N$^+$	34.65%	32.02%	26.42%	12.80%
NTM	63.18%	58.04%	53.04%	32.65%
NTM$^+$	**63.72%**	**60.77%**	**54.21%**	**33.90%**

Table 2: Bag-of-token model vs. Neural text model.

Model	NDCG@3	NDCG@5	NDCG@10	Pair Acc.
NTM	63.18%	58.04%	53.04%	32.65%
NTM/Pair Loss	**69.36%**	**66.36%**	**59.69%**	**35.67%**
NRM/Pair Loss	45.98%	46.37%	42.94%	32.41%

Table 3: Performance of neural text model vs. neural Retrieval Model

Model	NDCG@3	NDCG@3 Weighted	NDCG@5	NDCG@5 Weighted	NDCG@10	NDCG@10 Weighted
NTM	63.18%	125.87%	58.04%	128.44%	53.04%	114.37%
NTM/Pair Loss	69.36%	88.93%	69.36%	91.74%	59.69%	87.94%

Table 4: Performance of head query vs. tail query

Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*. 265–283.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *TACL* 5 (2017), 135–146.

[4] Eliot Brenner, Jun Zhao, Aliasgar Kutiyanawala, and Zheng Yan. 2018. End-to-End Neural Ranking for eCommerce Product Search: an application of task models and textual embeddings. In *eCom*.

[5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018).

[6] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

[8] Huizhong Duan, ChengXiang Zhai, Jinxing Cheng, and Abhishek Gattani. 2013. Supporting Keyword Search in Product Database: A Probabilistic Approach. *PVLDB* 6, 14 (2013), 1786–1797.

[9] Susan T. Dumais. 2016. Personalized Search: Potential and Pitfalls. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. 689.

[10] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Mix'n Match: Integrating Text Matching and Product Substitutability within Product Search. In *CIKM*.

[11] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*. 2333–2338.

[12] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 448–456.

[13] Mohit Iyyer, Varun Manjunatha, Jordan L. Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *ACL*. 1681–1691.

[14] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).

[16] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1746–1751.

[17] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer.

[18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[20] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval* 13, 1 (2018), 1–126.

[21] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. 807–814.

[22] Rama Kumar Pasumarthi, Xuanhui Wang, Cheng Li, Sebastian Bruch, Michael Bendersky, Marc Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. 2018. TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. *CoRR* abs/1812.00073 (2018).

[23] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[24] Ruslan Salakhutdinov and Geoffrey E. Hinton. 2009. Semantic hashing. *Int. J. Approx. Reasoning* 50, 7 (2009), 969–978.

[25] Gerard Salton, A. Wong, and Chung-Shu Yang. 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM* 18, 11 (1975), 613–620.

[26] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On Application of Learning to Rank for E-Commerce Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 475–484.

[27] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*. 101–110.

[28] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[29] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning Clicks into Purchases: Revenue Optimization for Product Search in E-Commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. 365–374.

[30] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 55–64.

[31] Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 1271–1279.

An Anatomy of a Lie:

Discourse Patterns in Customer Complaints Deception Dataset

Dina Pisarevskaya
Institute for Systems Analysis
FRC CSC RAS
Moscow, Russia,
dinabpr@gmail.com

Boris Galitsky
Oracle Corp
Redwood Shores CA, USA,
boris.galitsky@oracle.com

Jay Taylor
Oracle Corp
Redwood Shores CA, USA,
jay.taylor@oracle.com

Andrey Ozerov
Oracle Corp
Redwood Shores CA, USA, andrey.ozerov@oracle.com

ABSTRACT

Automated detection of text with misrepresentations such as fake reviews is an important task for online reputation management. The dataset of customer complaints - emotionally charged texts which are very similar to reviews and include descriptions of problems customers experienced with certain businesses – is presented. It contains 2746 complaints about banks and provides clear ground truth, based on available factual knowledge about the financial domain. Among them, 400 texts were manually tagged. Initial experiments were performed in order to explore the links between implicit cues of the rhetoric structure of texts and the validity of arguments, and also how truthful/deceptive are these texts.

CCS CONCEPTS

• Information systems → Information retrieval → Document representation → Content analysis and feature selection • Information systems → Information retrieval → Retrieval tasks and goals → Clustering and classification; Document filtering; Information extraction

KEYWORDS

Customer Complaints, Rhetorical Structure Theory, Discourse Analysis, Deception Detection, Fake Reviews

ACM Reference format:

Dina Pisarevskaya, Boris Galitsky, Jay Taylor and Andrey Ozerov. 2019. An Anatomy of a Lie:: Discourse Patterns in Customer Complaints Deception Dataset. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316468

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316468

1 Introduction

It has been discovered that a lot of forms of human intellectual and communication activity are associated with certain discourse structures. Rhetorical Structure Theory (RST) [8] is a good means to express correlation between such form of activity and its representation in how associated thoughts are organized in text. Rhetorical Structure Theory present a hierarchical, connected structure of a text as a Discourse Tree, with rhetorical relations between the parts of it. The smallest text spans are called elementary discourse units (EDUs). In communicative discourse trees (CDTs), the labels for communicative actions (CAs) (VerbNet expressions for verbs) are added to the discourse tree edges to show which speech acts are attached to which rhetoric relations; this structure helps to understand argumentation [5, 26].

Argumentation needs a certain combination of rhetorical relations of *Elaboration*, *Contrast*, *Cause* and *Attribution* [18]. Persuasiveness relies on certain structures linking *Elaboration*, *Attribution* and *Condition* [19]. Explanation needs to rely on certain chains of *Elaboration* relations plus *Explanation* and *Cause*, and a rhetorical agreement between a question and an answer is based on certain mappings between the rhetorical relations of *Contrast*, *Cause*, *Attribution* and *Condition* between the former and the latter [23, 27]. Discourse trees turned out to be helpful to form a dialogue and to build dialogue from text, to better understand the structure of texts.

In this paper, we study rhetoric structure correlated with certain forms of verbal activity, namely we focus on deception in texts such as reviews and complaints. Automated detection of fake reviews is important for online reputation management tasks. Natural Language Processing tools, that could distinguish truthful and reliable reviews from deceptive reviews, could be important for a broad spectrum of applications of recommendation and security systems, for wide range of products and services. Research on automated deception detection in written texts is focused on classifying if a narrative is truthful or deceptive. The main difficulty is to detect deception where factual knowledge is not available to a degree sufficient to

computationally establish the truth. This situation is typical in the every day life in the real world, from intuitive choice of product based on reviews to judges' verdicts: it is impossible to establish the truth based on known facts so decision are based on implicit cues such as the way people explain what they have done and provide arguments.

Detecting misrepresentation in writing, it is impossible to differentiate between different categories of writers. Professional writers are frequently good at misrepresenting, and they do not include cues for what might be a lie. Conversely, a content written by non-professional writers is often authentic in how it indicates the thought patterns of the writer where the traces of a lie and hints for how it is motivated can be found. Here, a corpora with defined ground truth are needed for classification tasks solving and exploring the links between implicit cues of rhetoric structure of texts and how truthful/deceptive are these texts [25].

2 Example of Misrepresentation in User-Generated Content

Regarding possible misrepresentation in a user-generated content, the following example can be provided:

> "I have accounts with them for almost 10 years, I hated it their customer service! Worst one ever. I don't know what's their problems, I'm not recommending their services and banking to anybody, I stopped using their credit cards already! The only reason I can't close my accounts with them, it could drop my credit score. I will not close my credit cards, but I'm not definitely using them so they can't make money from on us! I just had conversation with a supervisor from California called Steve he and his representative didn't even understand my situation, which was not common at all, basically didn't want to help me!"

The author of this complaint does not provide a single argument backing up his claim. And the author's statement that his credit history can be negatively affected by his closing an account is a misrepresentation.

We show the text split into elementary discourse units as done by discourse parser [20]. What do we see in the discourse tree for this text? We show important (non-default) rhetorical relations in bold and highlight with italics the verbs with the role of communicative actions which are important addition to the rhetorical relations.

Algorithm 1: A communicative discourse tree for the user-generated text example

elaboration (LeftToRight)
  elaboration (LeftToRight)
    attribution (LeftToRight)
      TEXT:I have accounts with them for almost 10 years ,
      TEXT:I *hated* it their customer service !
    TEXT:Worst one ever .
  elaboration (LeftToRight)
    elaboration (LeftToRight)
      explanation (LeftToRight)
        attribution (LeftToRight)
          cause (LeftToRight)
            attribution (RightToLeft)
              TEXT:I do not know
              TEXT:what is their problems ,
            TEXT:I 'm not *recommending* their services and banking to anybody ,
            TEXT:I *stopped using* their credit cards already !
          attribution (RightToLeft)
            TEXT:The only reason I can not close my accounts with them ,
            TEXT:it could drop my credit score .
        contrast (RightToLeft)
          TEXT:I will not close my credit cards ,
          enablement (LeftToRight)
            TEXT:but I 'm not definitely using them
            TEXT:so they can not make money from on us !
      elaboration (LeftToRight)
        TEXT:I just had conversation
        same-unit
          elaboration (LeftToRight)
            TEXT:with a supervisor from California called Steve , he and his representative did not even understand my situation ,
            TEXT:which was not common at all ,
          TEXT:basically did not want to help me !

There is an unusual chain of rhetorical relations explanation-attribution-cause-attribution-attribution which is a suspicious explanation pattern on its own. Unsurprisingly, the atom statement for the last attribution (which is the basis of this explanation, highlighted with underlined italics in Ex. 1) turns out to be false.

3 Example of Misrepresentation in Professional Writing

In our first example, the objective of the author is to attack a claim that the Syrian government used chemical weapon in the spring of 2018 (Fig. 1). An acceptable proof would be to share a certain observation, associated from the standpoint of peers, with the absence of a chemical attack. For example, if it is possible to demonstrate that the time of the alleged chemical attack coincided with the time of a very strong rain, that would be a convincing way to attack this claim. However, since no such observation was identified, the source, Russia Today, resorted to plotting a complex mental states expressing how the claim was communicated, which agents reacted which way for this communication. It is rather hard to verify most statements about the mental states of involved parties. We show the text split into EDUs as done by the discourse parser:

[Whatever the Douma residents ,][who had first-hand experience of the shooting of the water][dousing after chemical attack video ,][have to say ,][their words simply do not fit into the narrative][allowed in the West ,][analysts told RT .] [Footage of screaming bewildered civilians and children][being doused

with water ,][presumably to decontaminate them ,][was a key part in convincing Western audiences][that a chemical attack happened in Douma .] [Russia brought the people][seen in the video][to Brussels ,][where they told anyone][interested in listening][that the scene was staged .] [Their testimonies , however , were swiftly branded as bizarre and underwhelming and even an obscene masquerade][staged by Russians .] [They refuse to see this as evidence ,][obviously pending][what the OPCW team is going to come up with in Douma], [Middle East expert Ammar Waqqaf said in an interview with RT .] [The alleged chemical incident ,][without any investigation , has already become a solid fact in the West ,][which the US , Britain and France based their retaliatory strike on .]

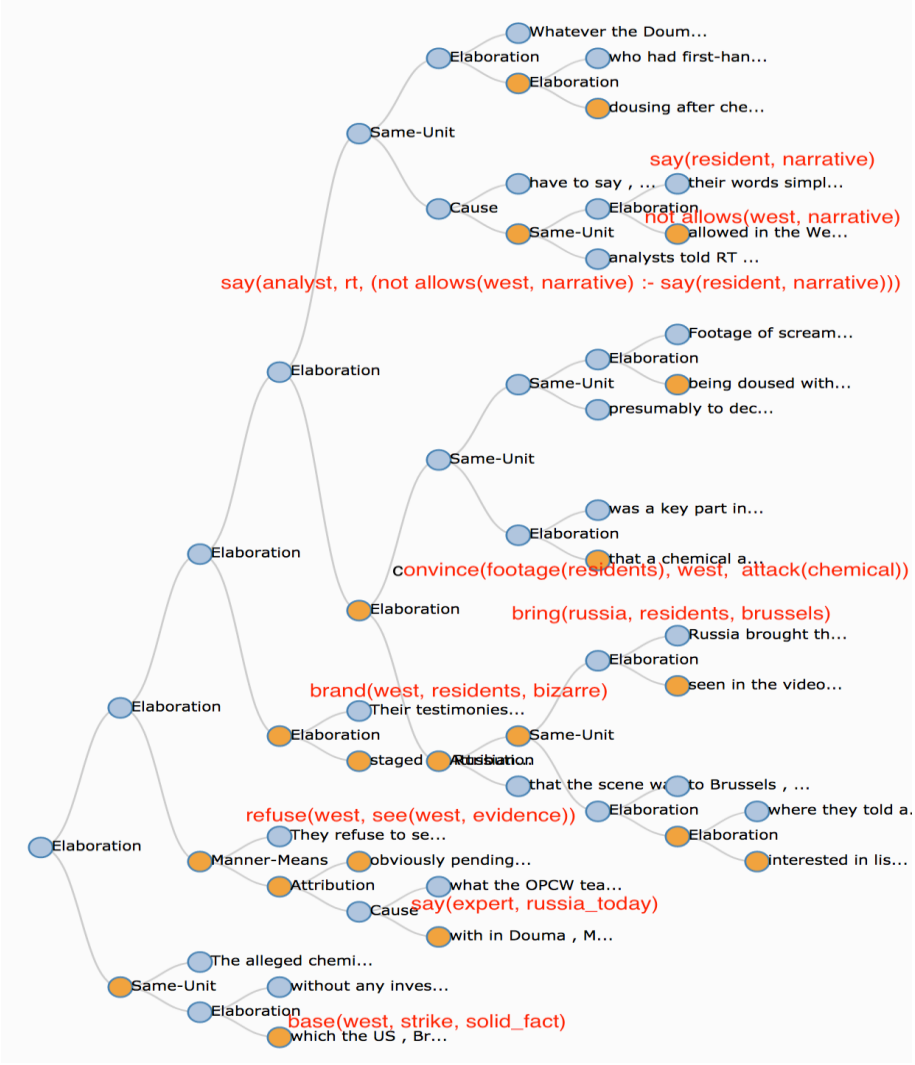


Figure 1: A DT for the chemical attack claim. An author attempts to substitute a desired valid argumentation chain by a fairly sophisticated mental states expressed by communicative actions.

This article (Fig. 1) does not really find counter-evidence for the claim of the chemical attack it attempts to defeat. Instead, the text says that the opponents are not interested in observing this counter-evidence. The main statement of this article is that a certain agent "disallows" a particular kind of evidence attacking the main claim, rather than providing and backing up this evidence. Instead of defeating a chemical attack claim, the article builds a complex mental states conflict between the residents, Russian agents taking them to Brussels, the West and a Middle East expert.

4 Background and Related Works on Deception Datasets

Deceptive product reviews can be referred to as deceptive opinion spam: fictitious opinions that have been deliberately

written to sound authentic, in order to deceive the reader [13]. Spammers write fake reviews to promote or demote target products. They are deliberately written to sound authentic, and it is difficult to recognize them manually: human average accuracy is merely 57.3% [13].

Automated deception detection for reviews faces the lack of gold standard corpora with verified examples of deceptive uses of language. Besides this, intentionally written (e.g. by crowdsourcing) texts are distinct from genuinely produced texts. Hence, such artificial texts classified as deceptive by human annotators are not necessarily totally deceptive.

The release of two gold-standard datasets (available at http://myleott.com/) allowed for applying supervised learning methods, taking stylistic, syntactic and lexical features into consideration [12, 13, 2, 3]. Hotels reviews were chosen for the datasets, because it was suggested that deception rates among travel reviews is reasonably small. The latter dataset includes, among other reviews, crowdsourced generation of deceptive reviews. It contains 400 truthful positive reviews from TripAdvisor; 400 deceptive positive reviews from Mechanical Turk; 400 truthful negative reviews from reviews websites; 400 deceptive negative reviews from Mechanical Turk.

Later researchers tried to overcome the lack of large realistic datasets on different topics and domains. For example, Yao et al. [17] apply a data collection method based on social network analysis to quickly identify deceptive and truthful online reviews from Amazon. The dataset contains more than 10,000 deceptive reviews in diverse product domains.

The problem of the mentioned above gold standard datasets is that the fake reviews were not taken from genuinely written ordinary reviews and manually classified as fake. Instead, they were written on demand by the Amazon Mechanical Turk workers, hence they are not indicative of deception [10]. However, they are accepted as gold standard datasets for this research field. Rules used in [13] to create ground truth datasets were also used in later projects, such as in [6].

The real-life Amazon dataset [7] contains reviews from Amazon.com (crawled in 2006). It is large and covers a very wide range of products. It was used, for example, in Sun et al. [16], namely, three domains: Consumer Electronics, Software, and Sports. The metadata in this dataset provides only helpfulness votes of the reviews.

In cases where there was no certain knowledge of the ground truth, different ways to collect reviews corpora, relying on other features, were used. For example, in [4] the DeRev corpus of books reviews, originally posted on Amazon, was collected using definite pre-defined deception clues. Book reviews in the corpus are marked as clearly fake, possibly fake, and possibly genuine. The corpus is constituted by 6,819 instances whose 236 were labeled with the higher degree of confidence and are considered as the gold standard.

In [14], two publicly available Yelp datasets were presented. They are labeled with respect to the Yelp's classification in recommended and not recommended reviews. Mukherjee et al. [9] found that the Yelp spam filter primarily relies on linguistic, behavioral, and social networking features. Classification provided by Yelp has been also used in many previous works before as a ground truth, where recommended reviews correspond to genuine reviews, and not recommended reviews correspond to fake ones, so these labels can be trusted. The YelpNYC dataset contains reviews of restaurants located in New York City (359,052 reviews; 10,27% are fake); the Zip dataset is larger, since it contains businesses located in contiguous regions of the U.S. (608,598 reviews; 13,22% are fake).

Big Amazon dataset is annotated with compliant/non-compliant labels. It has many different topics: from electronics and books to office products (https://s3.amazonaws.com/amazonreviewspds/readme.html). It contains labels about star rating, helpful vote, total votes, verified purchase, that could be used for making decisions.

Hence, the existing recent datasets rely on external factors provided by their source, such as review's rating, number of votes, social networking features of review's author, metadata features etc. They are not annotated manually. So, despite the presence of different corpora, lack of corpora with exact ground truth can be understood as a bottleneck in deception detection of online texts.

5 Dataset Description

We introduce the dataset of customer complaints – emotionally charged texts which include descriptions of problems they experienced with certain businesses. The dataset is freely available [24].

Raw complaints were collected from PlanetFeedback.com for a number of banks submitted in 2006-2010. The dataset consists of 2,746 complaints totally. 400 complaints were manually tagged with respect to the parameters related to argumentation and validity of text: perceived complaint validity, argumentation validity, presence of specific argumentation patterns, and detectable misrepresentation. Here, validity of information is connected with validity of arguments. The dataset contains texts with direct truth confirmation based on manual annotation. It contains authentic data: both truthful and deceptive reviews were taken from spontaneously written customers' texts. Among the manually annotated 400 complaints, 163 are invalid and 237 are valid.

The initial set of 80 complaints was tagged by the authors of the paper as experts. After that, three annotators worked with this dataset, having a set of definitions and applying them. Then precision and recall were measured by matching the tags done by the authors as the 'gold standard', after that the set of definitions was edited and elaborated. In the further work, the Krippendorff's alpha measure (for three annotators) was applied as inter-annotator agreement measurement, and it exceeds 80%. As it is possible to know, retrospectively and based on facts, the established ground truth, we suggest that the annotators can find out, with high confidence, what information in texts is deceptive. So the dataset would provide ground truth.

The rest 2,346 complains were auto-tagged based on the model trained on this 400 set. After that they have also been partially manually evaluated, so that the accuracy of auto tagging exceeds 75%.

Our dataset includes more complaints with intense argumentation in comparison with other argument mining datasets, such as [15, 1, 11]. For a given topic such as insufficient funds fee, this dataset provides many distinct ways of argumentation that this fee is unfair. Authors attempt to provide as strong argumentation as possible to back up their claims and strengthen their case.

If a complaint is not truthful it is usually invalid: either a customer complains out of a bad mood or wants to get compensation. However, if the complaint is truthful, it can still easily be invalid, especially when arguments are flawed. When an untruthful complaint has valid argumentation patterns, it is hard for an annotator to properly assign it as valid or invalid, without the guidelines. So, according to the guidelines for the manual tagging of the dataset, a complaint was considered as valid if a judge believed that the main complaint claim is truthful under the assumption that a complainant is making truthful statement. Valid complaint needs to include proper discourse and acceptable argumentation patterns.

Following this approach, a complaint is marked as truthful if a judge cannot defeat it, using commonsense knowledge, available factual knowledge about a domain or implicit, indirect cues. Inconsistencies detected by a judge also indicate that the complaint author is deceiving. Mentioning multiple unusual, very rarely occurring claims also indicate that the complaint author is deceiving. The judge does not have to be able to prove that the complainant is lying: judge's intuition is sufficient to tag a complaint as untruthful. We suggest that one can provide a valid argumentation and also provide a false statement in a single sentence: "Rule is like this <correct rule> and I followed it, making <false statement>". Conversely, one can be truthful but provide an invalid argumentation pattern "I set this account for direct deposit and sent a check out of it <truthful statement>, as my HR manager suggested <should not have followed advice from not a specialist in banking>". Therefore validity (of argumentation patterns) and truthfulness are correlated.

Furthermore, customer complaints have much more significance for well-being of customers in comparison with customer reviews. Therefore, tagged customer complaints have much more importance associated with truth/deception than customer reviews. Since reviews are associated with opinions which can be random and complaints with customers doing their best to achieve their goals, both the truth and a lie is much more meaningful and serious in comparison with review datasets.

Complaints usually have a simple motivational structure, are written with a fixed purpose. Most complainants face a strong deviation between what they expected from a service, what they received and how it was communicated. Most complaint authors report incompetence, flawed policies, ignorance, indifference to customer needs from the customer service personnel. The authors are frequently exhausted communicative means available to them, confused, seeking recommendation from other users and advise others on avoiding particular financial service. The focus of a complaint is a proof that the proponent is right and the opponent is wrong, resolution proposal and a desired outcome.

Complaints reveal shady practice of banks during the financial crisis of 2007, such as manipulating an order of transactions to charge a highest possible amount of non-sufficient fund fees. Moreover, the most frequent topic is about banks attempts to communicate this practice as a necessity to process a wide amount of checks. That's why the dataset collection is based on complaints of 2007.

Multiple argumentation patterns are used in complaints.

1. Deviation from what has happened from what was expected, according to common sense (most frequent). This pattern covers both valid and invalid argumentation (a valid pattern).
2. The second argumentation patterns cites the difference between what has been promised (advertised, communicated) and what has been received or actually occurred. It also mentions that the opponent does not play by the rules (valid).
3. A high number of complaints are explicitly saying that bank representatives are lying. Lying includes inconsistencies between the information provided by different bank agents, factual misrepresentation and careless promises (valid).
4. Complaints arise due to rudeness of bank agents and customer service personnel. Customers cite rudeness in both cases, when the opponent point is valid or not (and complaint and argumentation validity is tagged accordingly).
5. Complainants cite their needs as reasons bank should behave in certain ways. A popular argument is that since the government via taxpayers bailed out the banks, they should now favor the customers (invalid).

6 Communicative Discourse Trees to Represent Truthfulness in Text

Starting from the autumn of 2015, we became interested in the controversy about Theranos, the healthcare company that hoped to make a revolution in blood tests. Some sources including the *Wall Street Journal* started claiming that the company's conduct was fraudulent. The claims were made based on the whistleblowing of employees who left Theranos. At some point FDA got involved, and as the case develops, we were improving our deception detection techniques while keeping an eye on Theranos' story. As we scraped discussions about Theranos back in 2016 from the website, the audience believed that the case was initiated by Theranos competitors who felt jealous about the proposed efficiency of the blood test technique promised by Theranos. However, our analysis showed that Theranos was misrepresenting and our findings supported the criminal case against Theranos, which led to the massive fraud verdict. SEC says that Theranos CEO Elizabeth Holmes raised more than $700 million from investors "through an elaborate, years-long fraud" in which she exaggerated or made false statements about the company's technology and finances.

We now build an example of communicative discourse tree (CDT) for the Theranos attack on Wall Street Journal (WSJ)

acquisition (Fig. 2): "*It is not unusual for disgruntled and terminated employees in the heavily regulated health care industry to file complaints in an effort to retaliate against employers for termination of employment. Regulatory agencies have a process for evaluating complaints, many of which are not substantiated. Theranos trusts its regulators to properly investigate any complaints.*"

To show the structure of a deception, discourse relations are necessary but insufficient, and speech acts are necessary but insufficient as well. For the paragraph above, we need to know the discourse structure of interactions between agents, and what kinds of interactions they are.

file(employee, complaint) is elaborated by retaliate(employee, employer), and

evaluate(regulation, complaints) is elaborated by *trust(Theranos, regulators, complainants).* Also, the top link in turn is elaborated by the bottom link. Once we involve the definitions of the verbs for these four communicative actions, the inconsistency is revealed.

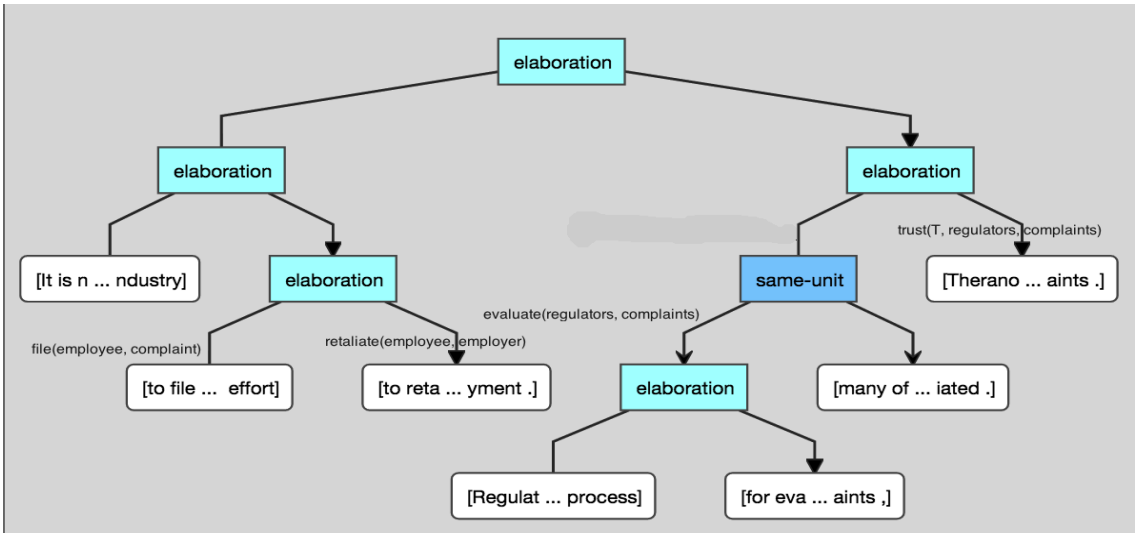


Figure 2: CAs as labels for rhetoric relations helps to identify a text apart from a heated discussion

From the commonsense reasoning standpoint, Theranos, the company, has two choices to confirm the argument that *his tests are valid*:

1) Conduct independent investigation, comparing their results with the peers, opening the data to the public, confirming that their analysis results are correct.
2) Defeat the argument by its opponent that their testing results are invalid, and providing support for the claim that their opponent is wrong.

Obviously, the former argument is much stronger, and we know, that usually the latter argument is chosen when the agent believes that the former argument is too hard to implement. On one hand, the reader might agree with Theranos that WSJ should have provided more evidence for its accusations against the company. On the other hand, the reader perhaps disliked the fact that Theranos selects the latter argument type (2) above, and therefore the company position is fairly weak.

The authors believe that Theranos' argument is weak because the company tries to refute the opponent's allegation concerning the complaints about Theranos's services from clients. We believe that Theranos' demand for evidence by inviting WSJ to disclose the sources and the nature of the complaints is weak. A claim is that a third-party (independent investigative agent) would be more reasonable and conclusive. However, some readers might believe that the company's argument (burden of proof evasion) is logical and valid.

7 Evaluation

In our evaluation we used the following pipelines:

Communicative Discourse Tree Construction. Just two RST parsers constructing discourse tree (DT) from paragraphs of text are available at the moment. We used the tool provided by [20, 21]. We then build CDT involving VerbNet.

Nearest Neighbor learning. To predict the label of the text, once the complete DT is built, one needs to compute its similarity with DTs for the positive class and verify that it is lower than similarity to the set of DTs for its negative class. Similarity between CDT's is defined by means of maximal common sub-DTs. Definitions of labeled graphs and domination relation on them used for construction of this operation can be found, e.g., in [22].

SVM Tree Kernel learning. A DT can be represented by a vector of integer counts of each sub-tree type (without taking into account its ancestors). For EDUs as labels for terminal nodes only the phrase structure is retained: we suppose to label the terminal nodes with the sequence of phrase types instead of parse tree fragments. For the evaluation purpose Tree Kernel builder tool [23] was used. After that, we applied the further set

of more complex experiments. For all texts, we use CDT-kernel learning approach. We combined Stanford NLP parsing, coreference resolution tool, entity extraction, CDT construction (based on automated discourse parser as in [20, 21]) and Tree Kernel builder into one system that is presented in [25].

For the initial and automatically derived datasets, we show (in bold) the accuracies of training row and testing, averaging through 5x cross-validation. For the bottom three datasets, we tested the same SVM Tree Kernel model trained on our dataset. We demonstrate its universality, showing its applicability to texts of various nature such as consumer reviews. For genuine reviews, only 380 cases of deception were detected which were false positives, assuming that review writers do not lie (Table 1).

Table 1: Datasets, evaluation settings, accuracies for deception detection initial experiments

Dataset	Deception	No deception	P	R	F1
Manually tagged complaints	163	237	91	85	**88**
			83	81	82
Automatically tagged based on initial classifier	1132	1615	78	75	**76**
			69	71	70
Genuine reviews	380	3420	83	100	91
Fake reviews	414	286	100	59	74

Here, for reviews datasets we use the dataset presented in [12,13], in order to compare two following assessment frameworks:

1) The framework [12] which is trained on consumer reviews and tested on similar dataset;
2) Our deception recognition framework which we train once on our own dataset and tested on texts of various kinds, such as reviews.

We achieved the performance between 74 and 91% which is not as high as in [12,13] but by the universal text classification system. Hence we expect it to detect deception in other text datasets with acceptable accuracy to assure a resultant decision support system is usable.

In conclusion, we mention that our dataset is in the initial stage and is still being developed. In the future studies, the whole complaints dataset should be manually annotated and used for model training and new experiments, it could possibly help in results improvement. We are also going to run our detector again the business communication dataset from the real word, for evaluation purposes. We also plan to run experiments with other machine leaning methods. We also suggest that precision improvement (reducing the number of false positives) is mostly important for deception detection task, so we will implement further steps to improve precision. After that, we could also develop different methods of the customer complaints dataset extension.

Both truthfulness and validity are recognized reasonably well which is a value for Customer Relation Management systems and could be useful in different e-commerce tasks that are based on online review analysis. The dataset could be used for different machine learning models training that could help detect if reviews or other online texts of similar genres are truthful or deceptive, based on their content features.

REFERENCES

[1] Abbott, R., Ecker, B., Anand, P., and Walker, M.A. (2016). "Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it". Language Resources and Evaluation Conference.

[2] Feng, S., Banerjee, R., and Choi, Y. (2012a). "Syntactic stylometry for deception detection". ACL 12, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: 171-175.

[3] Feng, S., Xing, L., Gogar, A., and Choi, Y. (2012b). "Distributional footprints of deceptive product reviews". Sixth International AAAI Conference on Weblogs and Social Media. 98-105.

[4] Fornaciari, T. and Poesio, M. (2014). "Identifying fake Amazon reviews as learning from crowds". Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics: 279-287.

[5] Galitsky, B., Ilvovsky, D., and Pisarevskaya, D. (2018). "Argumentation in text: Discourse structure matters". CICLing 2018 (unpublished).

[6] Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X.-L., and Li, G. (2016). "Deceptive review spam detection via exploiting task relatedness and unlabeled data". Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: 1817-1826.

[7] Jindal, N. and Liu B. (2008). "Opinion spam and analysis". Proceedings of International Conference on Web Search and Data Mining (WSDM-2008).

[8] Mann, W.C. and Thompson S.A. (1987). "Rhetorical Structure Theory: Description and Construction of Text Structures". Springer Netherlands, Dordrecht.

[9] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013a). "What Yelp fake review filter might be doing?" Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

[10] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013b). "Fake review detection: classification and analysis of real and pseudo reviews". Tech. rep. uic-cs-2013-03. University of Illinois at Chicago.

[11] Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M., and Whittaker, S. (2015). "And that's a fact: Distinguishing factual and emotional argumentation in online dialogue". NAACL HLT 2015 2nd Workshop on Argumentation Mining.

[12] Ott, M., Cardie, C., and Hancock, J.T. (2013). "Negative deceptive opinion spam". NAACLHLT 2013, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 497-501.

[13] Ott, M., Choi, Y., Cardie, C., and Hancock, J.T. (2011). "Finding deceptive opinion spam by any stretch of the imagination". Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1: 309-319.

[14] Rayana, S. and Akoglu, L. (2015). "Collective opinion spam detection: Bridging review networks and metadata". Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM: 985-994.

[15] Stab, S. and Gurevych, I. (2017). "Recognizing insufficiently supported arguments in argumentative essays". Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017): 980-990.

[16] Sun, C., Du, Q., and Tian, G. (2016). "Exploiting product related review features for fake review detection". Mathematical Problems in Engineering.

[17] Yao, W., Dai, Z., Huang, R., and Caverlee, J. (2017). "Online deception detection refueled by real world data collection". Proceedings of Recent Advances in Natural Language Processing: 793-802.

[18] Grasso, F. (2003). "Characterizing Rhetoric Argumentation". PhD Thesis Heriot-Watt University.

[19] Galitsky, B., Ilvovsky, D., and Makhalova, T. (2019). "Enabling a Bot with Understanding Argumentation and Providing Arguments". Developing Enterprise Chatbots, Springer Switzerland.

[20] Joty, S., Carenini, G., Ng, R.T., and Mehdad, Y. (2013). "Combining intra-and multisentential rhetorical parsing for document-level discourse analysis". ACL (1): 486-496.

[21] Surdeanu, M., Hicks, T., and Valenzuela-Escarcega, M.A. (2015). "Two Practical Rhetorical Structure Theory Parsers". NAACL HLT.

[22] Ganter, B. and Kuznetsov, S.O. (2001). “Pattern Structures and Their Projections”. ICCS, Lecture Notes in Artificial Intelligence (Springer), Vol. 2120: 129-142.
[23] Galitsky, B. (2019). “Rhetorical Agreement: Maintaining Cohesive Conversations”. Developing Enterprise Chatbots, Springer – Cham, Switzerland.
[24] Github (2019). Ultimate Deception Dataset https://github.com/bgalitsky/relevance-based-on-parse-trees/blob/master/examples/ultimateDeceptionAutoTagged.csv.zip.
[25] Galitsky, B. (2019). “A Content Management System for Chatbots”. Developing Enterprise Chatbots. Springer Switzerland.
[26] Galitsky, B. and Kuznetsov, S.O. (2008). “Learning communicative actions of conflicting human agents”. J. Exp. Theor. Artif. Intell, 20(4): 277–317.
[27] Galitsky, B., Ilvovsky, D., and Kuznetsov, S.O. (2018). “Detecting logical argumentation in text via communicative discourse tree". J. Exp. Theor. Artif. Intell., 30(5): 1-27.

The WDC Training Dataset and Gold Standard for Large-Scale Product Matching

Anna Primpeli
Data and Web Science Group
University of Mannheim
anna@informatik.uni-mannheim.de

Ralph Peeters
Data and Web Science Group
University of Mannheim
rpeeters@mail.uni-mannheim.de

Christian Bizer
Data and Web Science Group
University of Mannheim
chris@informatik.uni-mannheim.de

ABSTRACT

A current research question in the area of entity resolution (also called link discovery or duplicate detection) is whether and in which cases embeddings and deep neural network based matching methods outperform traditional symbolic matching methods. The problem with answering this question is that deep learning based matchers need large amounts of training data. The entity resolution benchmark datasets that are currently available to the public are too small to properly evaluate this new family of matching methods. The WDC Training Dataset for Large-Scale Product Matching fills this gap. The English language subset of the training dataset consists of 20 million pairs of offers referring to the same products. The offers were extracted from 43 thousand e-shops which provide schema.org annotations including some form of product ID such as a GTIN or MPN. We also created a gold standard by manually verifying 2200 pairs of offers belonging to four product categories. Using a subset of our training dataset together with this gold standard, we are able to publicly replicate the recent result of Mudgal et al. that embeddings and deep neural network based matching methods outperform traditional symbolic matching methods on less structured data.

CCS CONCEPTS

• **Information systems → Electronic commerce**.

KEYWORDS

entity resolution, product matching, schema.org annotations, deep matching, embeddings, evaluation data

ACM Reference Format:
Anna Primpeli, Ralph Peeters, and Christian Bizer. 2019. The WDC Training Dataset and Gold Standard for Large-Scale Product Matching. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316609

1 INTRODUCTION

As a result of the success of using embeddings and deep neural networks for image and speech recognition as well as natural language processing, the question whether these techniques also increase the performance of entity matching methods has recently moved into the research focus [12, 17, 18]. Current results by Mudgal et al. [12] suggest that deep learning techniques perform similar to traditional symbolic matching techniques on strongly structured data but outperform traditional techniques by a margin of 5% to 10% in F1 on less structured data such as product descriptions in e-commerce. The problem with these results is that they are not publicly reproducible as they have been achieved using large training datasets from a 'major retailer' [12] which are not available to the public.

Many e-shops have started to mark-up offers in HTML pages using schema.org annotations. In recent years, many of these e-shops have also started to annotate product identifiers within their pages, such as manufacturer part numbers (MPNs), global trade item numbers (GTINs), or stock keeping units (SKUs), a practice that was less common 5 years ago. These identifiers allow offers for the same product from multiple e-shops to be grouped into clusters. The *Web Data Commons* (WDC) project[1] monitors the adoption of schema.org annotations by analysing the CommonCrawl[2], a series of public web corpora each consisting of several billion HTML pages. Table 1 shows the number of pay level domains (PLDs) in the CommonCrawl that use product-related schema.org terms in 2017 compared to 2013. We see that the absolute numbers of websites, the richness of the descriptions, as well as the number of websites annotating product identifiers (lower part of the table) have all grown significantly.

This paper presents a large public training dataset for product matching which has been produced by extracting schema.org product descriptions that include identifiers from the CommonCrawl (November 2017). The training dataset consists of 26 million offers from 79 thousand websites. Using the identifiers and a specific cleansing workflow, the offers are grouped into 16 million clusters of offers referring to the same product. 1.1 million of these clusters have a size of three and larger, 413 thousand have a size of five and larger. The English language subset of this dataset consists of 16 million offers which are grouped into 10 million clusters. Out of these clusters, 625.7 thousand have a size of three and larger and 225 thousand have a size of five and larger. Only considering clusters of English offers having a minimum size of five and excluding clusters having a size larger than 80 which may introduce noise, 20.7 million positive training examples (pairs of matching product offers) and a maximum of 2.6 trillion negative training examples can be derived from the dataset. This means that the training dataset is several orders of magnitude larger than the largest training set for product matching that has been accessible to the public so far (see Section 6).

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316609

[1] http://www.webdatacommons.org/structureddata/
[2] http://commoncrawl.org/

Table 1: Adoption of product-related schema.org properties. The percent numbers refer to all websites using schema.org product markup.

Property	#PLDs 2013	#PLDs 2017	%PLDs 2013	%PLDs 2017
s:Product/name	50,536	535,625	89.62%	92.11%
s:Offer/price	33,509	462,444	59.42%	79.53%
s:Product/offers	33,090	462,233	58.68%	79.49%
s:Offer/priceCurrency	14,704	430,556	26.08%	74.04%
s:Product/image	34,921	419,391	61.93%	72.11%
s:Product/description	38,037	377,639	67.46%	64.94%
s:Offer/availability	21,789	337,876	38.64%	58.11%
s:Product/url	11,937	263,720	21.17%	45.35%
s:Product/brand	5,880	73,934	10.43%	12.71%
s:Product/productID	7,392	35,211	10.90%	6.05%
s:Product/sku	1,323	126,696	1.95%	21.78%
s:Product/mpn	484	8,161	0.71%	1.40%
s:Product/gtin13	276	5,467	0.41%	0.94%
s:Product/identifier	160	538	0.24%	0.09%
s:Product/gtin8	0	257	0.00%	0.04%
s:Product/gtin12	0	577	0.00%	0.09%
s:Product/gtin14	0	722	0.00%	0.12%

In addition to the training dataset, we build a gold standard for evaluating matching methods by manually verifying that 2200 pairs of offers refer or do not refer to the same products. The gold standard covers the product categories *computers, shoes, watches,* and *cameras*. Using both artefacts to publicly verify the results of Mudgal et al. [12], we find that embeddings and deep learning based methods outperform traditional symbolic matching methods by 5% to 11% in F1 on our gold standard.

This paper is structured as follows: Section 2 describes the data cleansing procedure that was used to create the training dataset. Section 3 profiles the cluster structure, the number of offers per product category, and the density of the schema.org attributes in the training dataset. We describe the creation and profile of the gold standard in Section 4. Section 5 presents the results using the datasets to train and evaluate various baseline matchers as well as the embeddings and deep learning based matchers proposed by Mudgal et al. Finally, Section 6 compares the training dataset and gold standard to existing evaluation datasets. The training set and the gold standard are provided for public download on the Web Data Commons website[3] which also provides additional statistics about both.

2 TRAINING DATASET CREATION

We use the Web Data Commons schema.org/Product data corpus November 2017[4], containing 809 million schema:Product and schema:Offer entities, as starting point for the creation of the training set. This section describes the cleansing procedure that we apply to derive the training set from the corpus. First, we focus on detecting offers having annotated identifiers and develop strategies to overcome syntactic errors in the annotations [7, 11]. Next, we group the offers using the identifiers into clusters. Finally, we extract key/value pairs from the HTML tables in the pages containing product specifications and categorize the offers into 26 product categories.

Table 2: Value overlap of different ID properties

Value overlap	gtin8	gtin12	gtin13	gtin14
sku	2,065	18,682	103,736	16,897
productID	2,966	13,854	198,847	10,192
identifier	122	3,751	17,732	837
Overlap #	5,153	36,287	320,315	27,926
Overlap %	6.75%	15.81%	11.21%	11.35%

Selection of schema.org identifier related properties. According to the schema.org/Product[5] and schema.org/Offer[6] property definitions, the terms *gtin* and *mpn* should be used to annotate global-scoped identifiers. Vendor specific identifiers should be marked-up as *sku*, while *productID* and *identifier* can be used either to markup vendor-specific or global identifiers. However, we observe that the identifier related terms are used inconsistently in many cases, as vendor-scoped terms are often used to annotate global-scoped identifier values. More specifically more than 19% of the distinct global identifier values are annotated using the property *sku* while the properties *identifier* and *productID* are also often used for the same purpose as shown in Table 2. Based on this observation, we consider all *schema.org/Product* and *schema.org/Offer* entities that include some form of identifying information using the properties *gtin8, gtin12, gtin13, gtin14, mpn, sku, identifier,* and *productID* for the creation of the training corpus.

Usage of non-existing schema.org terms. Similar to the observations of [11], we notice that 6% of the websites annotating product offers with identifier values use invalid schema.org terms. A frequent error pattern is the usage of non-existing schema.org types, such as *IndividualProduct/productID* or *ProductModel/sku*. Despite the wrong vocabulary usage such terms reveal identifying information for an offer and we do not want to ignore them upon the creation of the training set. We capture such offers by applying the following regular expression pattern on their predicates: $.*/(gtin8|gtin12|gtin13|gtin14|sku|mpn|identifier|productID)$. This results in 116 million offer entities being selected from the WDC schema.org/Product subcorpus.

Leveraging entity relations. Approximately 20% of the offers do not contain descriptive properties such as *schema.org/name* and *schema.org/description*. This originates in the annotation practice of describing a product and an offer for this product using two separate schema.org entities. The two entities are connected using properties such as *Product/offers* while the descriptive and identifying information is split between the two entities. We identify those entity relations and merge the descriptive information. This leads to a reduction of offers having no descriptive properties to less than 3%.

[3] http://www.webdatacommons.org/largescaleproductcorpus/
[4] http://www.webdatacommons.org/structureddata/2017-12/stats/schema_org_subsets.html
[5] https://schema.org/Product
[6] https://schema.org/Offer

Detection and removal of listing pages and advertisements. We want to include the comprehensive information about a product from its detail page into the corpus and not the summaries of this information often found on listing pages and in advertisements on other detail pages. For the detection of listing pages and advertisements a heuristic based on the following features is applied: amount of schema.org/Offer and schema.org/Product entities per webpage, variation of the length of the product descriptions, number of identifier values, and semantic connection to parent entities using the terms schema:RelatedTo and schema:SimilarTo. Our heuristic for identifying listing pages and advertisements achieves an F1 score of 94.8% on a manually annotated test set. This cleansing step removes 49% of the offer entities, leaving 58 million non listing or ad offer entities in the training set.

Filtering by identifier value length and occurrence. As next pre-processing step, the annotated identifier values are normalized by removing non-alphanumeric characters and common prefixes such as initial zero digits and identifier related strings like *ean, mpn, sku,* and *isbn*. Considering the length of global identifiers such as GTIN or ISBN numbers and the fact that short identifiers more likely introduce noise in the cluster creation phase, we filter out all offers having identifiers which are shorter than 8 characters. Additionally, offers whose id values completely consist of alphabetical characters are removed. Finally, we observe that a considerable amount of websites use the same identifier value to annotate all their offers, likely due to an error in the script generating the pages. We detect these websites and remove their offers from the training set. After these filtering steps 26 million offer entities remain in the training corpus.

Clusters creation. We group the remaining 26 million offers into 18 million clusters using their identifier values. It happens that single offers contain multiple alternative identifiers referring to the same product, e.g. GTIN8 and GTIN12, or GTIN12 and MPN. We use this information to merge clusters referring to the same product which results in a reduction of the number of clusters to 16,391,439. We also note that some websites include identifiers referring to product categories, such as UNSPSC numbers, in addition to identifiers referring to single products into the annotations. For detecting such cases, we examine the structure of the identifier co-occurrence graph within each cluster. We discover that vertices having a degree larger than 10 and a clustering coefficient of $C_i < 0.2$ tend to represent product categories rather than single products and we split the clusters accordingly. This leads to the creation of 199,139 additional clusters.

Specification tables detection and extraction. Product pages often contain specification tables describing the product in the form of key/value pairs. This structured product data is often very helpful for matching [6, 14]. Based on the work of [13, 16] on detecting specification tables in HTML pages, we apply a table detection heuristic considering the following HTML table attributes: ratio of alphanumeric characters, average number of columns per row, table children elements, image occurrences, maximum number of columns, maximum number of rows, and average length of text per row. Afterwards, we use the two column-heuristic from [16] to extract key/value pairs from tables. Evaluating our specification table detection heuristic on 455 manually annotated HTML tables, we find that the heuristic reaches an F1 of 78%.

Table 3: Distribution of offer entities and positive pairs per cluster size

Cluster size	# Clusters		# Positive Pairs	
	Full Set	English Set	Full Set	English Set
1	13,301,842	8,434,389	0	0
2	1,915,909	1,012,220	1,915,909	1,012,220
[3-4]	760,360	400,522	3,026,997	1,552,680
[5-10]	304,379	163,356	5,677,852	3,064,532
[11-20]	63,981	37,562	6,752,150	3,751,124
[21-30]	17,710	10,567	5,374,523	3,185,935
[31-40]	10,863	4,461	6,666,546	2,646,306
[41-50]	6,318	2,504	6,281,387	2,502,691
[51-60]	2,663	1,300	3,978,483	1,972,822
[61-70]	1,378	832	2,891,198	1,750,014
[71-80]	1,058	682	2,960,532	1,899,880
[>80]	4,978	3,999	137,488,518	113,935,797

Offer categorization. E-shops use a wide range of different categorization schemata to present their offers. In order to consistently categorize all offers into the same set of product categories, we apply transfer learning similar to [14]. Using a publicly available amazon.com dataset of product reviews and metadata[7], we build lexica containing terms and their TF-IDF scores for 26 product categories. For every offer in the training set we calculate a score for each category by considering the overlapping terms after TF-IDF weighting between the category lexicon and the schema.org properties *name, title, description*, and *brand* of the offer entity. The offer is assigned the category with the maximum score. In case the term overlap is minimal, the offer is assigned the category label "not found". Finally, we use majority voting among the offers of a cluster to assign cluster specific category labels. Thus, all offers belonging to one cluster are labeled with the same product category.

3 TRAINING DATASET PROFILING

This section analyses the structure of the WDC Training Dataset for Large-Scale Product Matching as well as the structure of its English-language subset. The English-language subset is created by selecting all offers from pages having the suffixes: com, net, co.uk, and org. The Full training set contains 26 million offer entities, deriving from 79 thousand websites, grouped into 16 million clusters. The English training set contains 16 million offer entities, deriving from 43 thousand websites, grouped into 10 million clusters. Table 3 shows the distribution of offer entities per cluster in the Full and English training sets as well as the amounts of positive pairs. We observe that small clusters (size one and two) account for 92% of the clusters in the Full training set. The reasons for the large fraction of small clusters are twofold: First, the long tail distribution of products on the Web. Second, the limited depth of the CommonCrawl as only a fraction of the pages of each website is crawled. However, the English training set contains over 600 thousand clusters having a size of three or larger.

Product Categories in the English Training Set.

[7] http://jmcauley.ucsd.edu/data/amazon/

Table 4 shows the distribution of clusters per product category as well as the clusters size distribution in the English training set. Considering clusters of size three and larger, the positive pairs for the categories *Shoes, Camera and Photo, Cell Phone and Accessories, Computers and Accessories,* and *Jewelry* are more than 250 thousand while for the categories *Office Product* and *Clothing* there are more than 1 million positive pairs.

Table 4: Distribution of product categories in the English training set

Category	% Clusters	# Clusters of Size			
		[3-4]	[5-10]	[11-20]	[>20]
Office	10.90	40,314	16,920	5,953	3,043
Jewelry	7.79	33,156	13,329	3,352	2,037
Clothing	6.80	49,085	30,384	3,285	1,866
Automotive	5.90	16,650	8,139	2,865	2,140
Beauty	5.78	27,636	10,568	2,115	1,070
Phones & Acc.	4.76	15,870	5,162	1,085	878
Home & Kitchen	4.68	24,429	7,414	1,247	538
Luggage	4.48	14,401	6,399	1,198	957
Tools	4.35	12,033	4,407	1,248	1,042
CDs & Vinyl	4.19	17,666	6,013	1,417	663
Shoes	4.11	16,603	7,590	1,335	721
Camera & Photo	3.47	14,583	5,408	1,423	935
Grocery	3.26	17,109	5,889	2,154	716
Computers & Acc.	3.20	11,614	5,411	2,308	2,862
not found	3.07	12,964	4,088	942	267
Digital Music	3.03	7,954	3,046	640	535
Other Electronics	2.83	11,649	4,412	977	427
Books	2.81	9,889	2,946	330	183
Video Games	2.62	8,256	3,419	938	779
Garden	2.43	4,898	1,764	475	366
Musical Instr.	2.31	5,182	1,684	550	486
Pet Supplies	2.15	7,605	2,974	620	440
Baby	1.71	5,509	1,894	458	254
Toys	1.19	3,120	1,016	258	189
Sports	0.75	3,460	1,234	314	372
Movies & TV	0.71	2,030	681	195	157
Health	0.70	6,857	1,165	189	113

Property distribution. Table 5 shows the distribution of the schema.org properties that are used to describe the offers in the Full and English training sets. In addition, we report the distribution of identifier related schema.org properties in both sets. Our specification table detection method finds at least one specification table in 24% of the HTML pages contained in the Full set and 17% of the pages of the English set. Using the key/value pair extraction heuristic described in Section 2, we are able to extract ten or more key/value pairs from 73% of the specification tables. Finally, 80% of the offers in both the Full and English training sets contain the *dcterms:title* property of the DCMI Type Vocabulary[8] which captures the content of the HTML <title> element.

[8]http://dublincore.org/documents/2008/01/14/dcmi-type-vocabulary/

Table 5: schema.org properties in the Training Dataset and Gold Standard

Property	Offers Full		Offers English		Gold	
	# (in k)	%	# (in k)	%	#	%
name	25,281	95.3	15,653	95.1	2,300	99.6
description	17,215	64.9	11,352	69.0	1,884	81.6
brand	9,313	35.1	5,645	34.3	767	33.2
image	5,785	21.8	4,348	26.4	407	17.6
price	3,335	12.5	1,977	12.0	301	13.0
priceCurr.	2,971	11.2	1,873	11.3	293	12.6
availability	1,180	4.4	716	4.3	170	7.3
manufact.	2,024	7.6	1,254	7.6	325	14.0
sku	11,475	43.2	7,239	44.0	747	32.3
mpn	4,611	17.3	3,167	19.2	1,504	65.1
productID	9,386	35.4	6,351	38.6	348	15.0
gtin8	452	1.7	167	1.0	130	5.6
gtin13	3,529	13.3	1,449	8.8	263	11.3
gtin12	300	1.1	261	1.5	16	0.6
gtin14	420	1.5	71	0.4	9	0.3
identifier	179	0.6	65	0.3%	5	0.2

4 WDC GOLD STANDARD FOR LARGE-SCALE PRODUCT MATCHING

Due to the general noisiness of web data as well as anomalies not resolved by our cleansing procedure, there is no guarantee that all offers in a cluster will always refer to the same product or that products in different clusters are always different. In order to allow matching methods to be evaluated on completely clean data, we create the WDC Gold Standard by manually verifying for a set of 2200 pairs of offers whether they refer to the same product or not.

The difficulty of a matching task as well as the suitable matching method for a task both depend on the structuredness of the data to be matched. Thus, we select two product categories containing less structured offers (watches and sneaker shoes) and two categories containing more structured offers (computers & accessories and camera & photo) and create one gold standard for each using offer pairs from our corpus. First, we identify the clusters belonging to the selected product categories. We select 75 related clusters preferring clusters having a large diversity among the offers' textual content and a minimum size of 5 offers. Large diversity in this context refers to offers for the same product that vary on the Jaccard string similarity of their titles and descriptions, thus leading to a selection of clusters that contain textually similar as well as less similar offers.

In order to select challenging pairs of offers for the manual verification, we apply the following procedure: From every selected cluster we pick one offer and exploit its textual content given by the dcterms:title, schema:name, schema:description, and specification table values. Similarly to [9] we use the Jaccard similarity metric and the offers' textual content to calculate the similarity scores between the picked offer and the offers of the same cluster (intra-cluster similarity scores) as well as the offers of different clusters from the full corpus(inter-cluster similarity scores). We select the intra-cluster offer pairs with the highest and lowest similarity scores and add them in the gold standard. In addition, we add one to

three inter-cluster offer pairs with the highest similarity score and three randomly chosen inter-cluster pairs in the gold standard. We manually verify that the selected pairs are really matches or non-matches by reading the textual content of the offers. If we discover that a pair is incorrectly labeled, we correct the label. Finally, we remove the manually annotated pairs from the training set.

The resulting gold standard datasets consist of 150 positive and 400 negative pairs for each category. The offers contained in the gold standard datasets originate from the following numbers of clusters for each category: 338 for Computers & Accessories, 231 for Camera & Photo, 269 for Watches and 186 for Sneakers. The two right-most columns in Table 5 describe the density of the properties of the offers contained in the gold standard.

5 BASELINE IDENTITY RESOLUTION EXPERIMENTS

This section presents a set of matching experiments conducted using the English training set and the WDC gold standard. The experiments are intended on the one hand to verify the utility of the WDC training set as well as the cleansing procedure that was used to create the training set. On the other hand, we use the training set and gold standard to publicly replicate the results of Mudgal et al. [12]. First, we perform an unsupervised bag-of-words experiment using TF-IDF and cosine similarity. Afterwards, we train various supervised models such as logistic regression, naive Bayes, LinearSVC, decision trees, and random forests using (i) binary word co-occurrence vectors and (ii) string similarity scores, automatically generated by the Magellan framework [8], as features. As neural network based matchers, we combine all network types implemented in the *deepmatcher* framework (e.g. RNNs, Attention, and Hybrid) with pre-trained and self-trained *fastText* embeddings.

We experiment with different subsets of the offer features title, description, brand, and specification table content. For the title feature we concatenate the textual values schema:name and dcterms:title. All identifier related properties (lower part of the Table 5) are removed from the offers. Due to resource limitations, we do not use the complete English training set for the supervised experiments but subsets of potentially interesting training examples (e.g. positive pairs from many different clusters and negative pairs from different clusters where both offers have a rather similar description). Table 6 gives an overview of the size (number of positive and negative pairs), diversity (number of origin clusters) and feature density. We abbreviate the features schema:name and dcterms:title with T, schema:description with D, schema:brand with B and specification tables with S.

The results of the experiments are summarized in Table 7. For each category, we report the best performing method/feature combinations. As expected the supervised methods outperform the unsupervised BOW approach significantly. More interestingly, the deep learning approaches using *fastText* embeddings are 5-11% better in F1 compared to the supervised methods using symbolic feature representations. This confirms the result of Mudgal et al. that deep learning based matching methods excel on tasks involving rather textual entity descriptions. More information about the exact configuration of all methods as well as the results of method/feature combinations of weaker performance are found on the project's website.

Table 6: Training set profiling

Category	# Pos. Pairs	# Neg. Pairs	Clusters	% Feat. Density		
				T	D	S
Computers	20,444	21,676	338	100	82	55
Cameras	7,539	9,093	231	100	61	4
Watches	5,449	8,819	269	100	48	4
Shoes	3,476	5,924	186	99	36	1

Table 7: Results of all experiments

Category	Classifier	Features	P	R	F1
Unsupervised Matching					
Computers	Cosine, TF-IDF	T+D+B	0.52	0.70	0.60
Cameras	Cosine, TF-IDF	T+D+B	0.52	0.83	0.64
Watches	Cosine, TF-IDF	T	0.45	0.89	0.60
Shoes	Cosine, TF-IDF	T	0.61	0.76	0.67
Supervised Matching - Symbolic Features					
Computers	LinearSVM	T+D	0.75	0.94	0.84
Cameras	LinearSVM	T+D+B+S	0.70	0.87	0.78
Watches	LinearSVM	T+D+B+S	0.74	0.91	0.81
Shoes	LinearSVM	T+D+B+S	0.72	0.95	0.82
Computers	RandomForest	T	0.72	0.92	0.81
Cameras	RandomForest	T+D+B+S	0.75	0.87	0.81
Watches	RandomForest	T+D+B	0.66	0.91	0.77
Shoes	RandomForest	T+D+B+S	0.67	0.95	0.79
Supervised Matching - Symbolic Features - Magellan					
Computers	RandomForest	T+D	0.59	0.79	0.67
Cameras	RandomForest	T+D+B+S	0.53	0.85	0.65
Watches	RandomForest	T+D+B+S	0.71	0.85	0.78
Shoes	RandomForest	T+D+B+S	0.71	0.95	0.81
Supervised Matching - Distributed Features - DeepMatcher					
Computers	RNN	T+D+B+S	0.84	0.96	**0.89**
Cameras	RNN	T+D+B+S	0.88	0.93	**0.90**
Watches	RNN	T+D+B+S	0.88	0.97	**0.92**
Shoes	RNN	T+D+B+S	0.88	0.97	**0.92**

6 RELATED WORK

This section compares the *WDC Training Dataset for Large-Scale Product Matching* with existing resources for the evaluation of entity resolution methods. There exist various evaluation datasets for the task of product matching. The two classic datasets in this area *Abt-Buy* and *Amazon-Google* were introduced by Köpcke and Rahm [10]. Gokhale et al. introduce another public product dataset *Walmart-Amazon* [5]. Mudgal et al. [12] use several large product datasets for evaluating their deep learning methods. Unfortunately, these datasets are not public. Various benchmark datasets have been introduced for the Instance Matching Track of the Ontology Alignment Evaluation Initiative (OAEI)[9] over the years. Daskalaki et al. give an overview of these datasets [4]. The 2017 OAEI Instance

[9] http://oaei.ontologymatching.org/

Table 8: Comparison of evaluation datasets for entity resolution

Dataset	Public	# Data Sources	# Positive Pairs
Walmart-Amazon [5]	yes	2	1,154
Amazon-Google [10]	yes	2	1,300
Abt-Buy [10]	yes	2	1,097
DBLP-ACM [10]	yes	2	2,224
DBLP-Scholar [10]	yes	2	5,347
DM-Clothing [12]	no	1	105,608
DM-Electronics [12]	no	1	98,401
DM-Home [12]	no	1	111,714
DM-Tools [12]	no	1	96,836
DM-Company [12]	yes	?	28,200
OAEI - SYNTHETIC [1]	yes	1	1,800
Citeceer - DBLP [2]	yes	2	558,787
Falcon - Songs [3]	yes	1	1,292,023
WDC - Product GS [15]	yes	32	1,500
WDC - LSPM	yes	79,126	40,582,671
WDC - LSPM English	yes	43,293	20,773,304

Matching Track used the evaluation datasets SYNTHETIC and DOREMUS [1]. A large citation dataset *Citeseer - DBLP* offering 550 thousand matches is provided in the Magellan Data Repository [2]. A large song dataset containing 1.2 million matching pairs has been used to evaluate Falcon [3]. As part of our previous work [15], we have published a gold standard for product data extraction and matching covering 32 different e-shops.

Table 8 compares the *WDC Training Dataset for Large-Scale Product Matching (WDC - LSPM)* to other evaluation datasets along the dimensions number of positive pairs (e.g. offers referring to the same product) as well as number of sources from which the data originates. The table shows that concerning the number of positive pairs WDC - LSPM is several orders of magnitude larger than the existing evaluation datasets in the area of product matching (public as well as proprietary datasets). Compared to the Falcon-Songs dataset, WDC - LSPM English is 17 times larger. Concerning the number of sources, WDC - LSPM English covers 43,293 sources while the existing datasets cover at most 32 sources.

7 CONCLUSION

The *WDC Training Dataset* nicely demonstrates the utility of the Semantic Web. Without the website owners putting semantic annotations into their HTML pages it would have been much harder, if not impossible, to extract product offers from 79 thousand e-shops. While the training set likely still contains some noise that the cleansing procedure did not remove, being able to achieve F1 scores around 0.90 in the experiments clearly demonstrates the utility of the training set. We hope that researchers working on entity resolution and e-commerce will consider the *WDC Training Dataset and Gold Standard* useful and we hope that both artefacts contribute to advance the understanding of the potentials of latent semantic representations and deep neural networks for the task of product matching.

REFERENCES

[1] Manel Achichi, Michelle Cheatham, et al. 2017. Results of the Ontology Alignment Evaluation Initiative 2017. In *Proceedings of OM 2017-12th ISWC workshop on ontology matching*. 61–113.

[2] Sanjib Das, AnHai Doan, et al. 2016. The Magellan Data Repository. https://sites.google.com/site/anhaidgroup/useful-stuff/data.

[3] Sanjib Das, Paul Suganthan G.C., et al. 2017. Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17)*. 1431–1446.

[4] Evangelia Daskalaki, Giorgos Flouris, et al. 2016. Instance Matching Benchmarks in the Era of Linked Data. *Journal of Web Semantics* 39 (2016), 1 – 14.

[5] Chaitanya Gokhale, Sanjib Das, et al. 2014. Corleone: Hands-off Crowdsourcing for Entity Matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD '14)*. 601–612.

[6] Anitha Kannan, Inmar E. Givoni, et al. 2011. Matching Unstructured Product Offers to Structured Product Specifications. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. 404–412.

[7] Elias Kärle, Anna Fensel, et al. 2016. Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema.org Usage in the Hotel Domain. In *Information and Communication Technologies in Tourism 2016*. Cham, 99–112.

[8] Pradap Konda, Jeff Naughton, and et al. 2016. Magellan: toward building entity matching management systems. *Proceedings of the VLDB Endowment* 9, 12 (2016), 1197–1208.

[9] Hanna Köpcke and Erhard Rahm. 2008. Training selection for tuning entity matching.. In *Proceedings of the 6th International Workshop on Quality in Databases and Management of Uncertain Data (QDB/MUD '08)*. 3–12.

[10] Hanna Köpcke, Andreas Thor, and Erhard Rahm. 2010. Evaluation of Entity Resolution Approaches on Real-world Match Problems. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 484–493.

[11] Robert Meusel and Heiko Paulheim. 2015. Heuristics for Fixing Common Errors in Deployed Schema.Org Microdata. In *Proceedings of the 12th European Semantic Web Conference on The Semantic Web. Latest Advances and New Domains - Volume 9088*. 152–168.

[12] Sidharth Mudgal, Han Li, et al. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*. 19–34.

[13] Petar Petrovski and Christian Bizer. 2017. Extracting Attribute-value Pairs from Product Specifications on the Web. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. 558–565.

[14] Petar Petrovski, Volha Bryl, and Christian Bizer. 2014. Integrating Product Data from Websites Offering Microdata Markup. In *Companion Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. 1299–1304.

[15] Petar Petrovski, Anna Primpeli, et al. 2017. The WDC Gold Standards for Product Feature Extraction and Product Matching. In *E-Commerce and Web Technologies*. 73–86.

[16] Disheng Qiu, Luciano Barbosa, et al. 2015. Dexter: Large-scale Discovery and Extraction of Product Specifications on the Web. *Proceedings of VLDB Endowment* 8, 13 (2015), 2194–2205.

[17] Petar Ristoski, Petar Petrovski, et al. 2018. A machine learning approach for product matching and categorization. *Semantic Web* 9, 5 (2018), 707–728.

[18] Kashif Shah, Selcuk Kopru, et al. 2018. Neural Network based Extreme Classification and Similarity Models for Product Matching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. 8–15.

Towards language-agnostic alignment of product titles and descriptions: a neural approach

Daniel Stein
eBay Inc., Aachen, Germany
danstein@ebay.com

Dimitar Shterionov
Dublin City University, Dublihn, Ireland
dimitar.shterionov@adaptcentre.ie

Andy Way
Dublin City University, Dublihn, Ireland
andy.way@adaptcentre.ie

ABSTRACT

The quality of e-Commerce services largely depends on the accessibility of product content as well as its completeness and correctness. Nowadays, many sellers target cross-country and cross-lingual markets via active or passive cross-border trade, fostering the desire for seamless user experiences. While machine translation (MT) is very helpful for crossing language barriers, automatically matching existing items for sale (e.g. the smartphone in front of me) to the same product (all smartphones of the same brand/type/colour/condition) can be challenging, especially because the seller's description can often be erroneous or incomplete. This task we refer to as item alignment in multilingual e-commerce catalogues. To facilitate this task, we develop a pipeline of tools for item classification based on cross-lingual text similarity, exploiting recurrent neural networks (RNNs) with and without pre-trained word-embeddings. Furthermore, we combine our language agnostic RNN classifiers with an in-domain MT system to further reduce the linguistic and stylistic differences between the investigated data, aiming to boost our performance. The quality of the methods as well as their training speed is compared on an in-domain data set for English–German products.

ACM Reference Format:
Daniel Stein, Dimitar Shterionov, and Andy Way. 2019. Towards language-agnostic alignment of product titles and descriptions: a neural approach. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316602

1 INTRODUCTION

In the field of natural language processing (NLP), recent advances in deep learning have led to neural methods surpassing traditional rule-based or statistical ones for various tasks. One important drawback of such methods is the demand for large amounts of high-quality training data. While the situation of freely available (multi-)lingual training material continues to improve, obtaining suitably tailored in-domain content continues to be a bottleneck for training realistic models.

In the e-Commerce domain, whenever large proportions of the content are user-generated, even aligning and evaluating entries (mono- or cross-lingual) becomes extra difficult. These entries often fall short in terms of linguistic quality, and many are incomplete or contain incorrectly labelled data. Due to cultural preferences, legal restrictions in between countries, or other limitations, cross-country meta data (e.g. an overarching catalogue tree) is also likely to differ.

The work we summarize in this paper aims to align (eBay) cross-language item titles[1] that belong to the same product, via text similarity approaches. With this work we aim to provide a basis for three higher-level tasks that are of interest to eBay: identifying in-domain comparable data, content synchronization and detecting erroneous/misaligned entries (e.g. items that share the same product code but are actually distinct). With real-time application and fast training cycles in mind, we implement several neural methods and compare their performance in terms of accuracy and speed on selected English to German catalogue entries.

The paper is organized as follows: in Section 2 we present an overview of existing methods that tackle similar problems; in Section 3 we provide further motivation to our work by elaborating on use-cases; in Section 4 we describe the data we used; in Section 5 we discuss the methods and methodology we undertook; our experiments and results are summarized in Section 6; we conclude and raise points for future work in Section 7.

2 RELATED WORK

The exponential growth of multilingual content on the web, including commercial content, has created the necessity of higher-level taxonomic organization. Prytkova et al. [18] assess the necessity for alignment of multilingual taxonomies and propose several methods, including a string-similarity method using the Wikipedia taxonomy as a translation or a mapping medium. Among others, Fu et al. [4] and Spohr et al. [23] argue that, in the context of cross-lingual ontology matching, the quality of the Machine Translation (MT) system used is of major importance. The work of Nikoulina et al. [15] investigates cross-lingual search in library catalogues using MT adapted with a corpus of bilingual queries. The work of Guha and Heger [5] and, more recently of Sloto et al. [22], present challenges and solutions for MT for an e-Commerce vendor. In one of our approaches we also exploit MT to translate the German catalogues into English, and then compute string (text) similarity between items from the two catalogues in order to identify which matching entries are correctly aligned and which are not.

To bridge the language barrier in the context of cross-lingual information retrieval (IR), Eigen-analysis has also been used. In [9]

[1] Within this work an *item* is viewed as a container that encapsulates data such as title, description, identifier, etc., about a specific product that is for sale; multiple items can represent the same product; items are organized in the leaves of the catalogue trees we investigate.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316602

Table 1: Aligned entries from the training material: sample English entry, one of the UPC-aligned German entries, and a human translation of the German title to better read the differences.

EN:	Schleich 13818 Hanoverian Foal Toy Figure, Brown, For Ages 3+
DE:	Schleich Hannoveraner Fohlen 13818
trans.:	*Schleich Hannoverian Foal 13818*
EN:	Siku 4066 Crown Big x 580 Forage Harvester Scale 1:3 2 NEW! °
DE:	SIKU Krone Big X 580 Maishäcksler (Spielware) NEU
trans.:	*SIKU Crown Big X 580 Forage Harvester (Toy) NEW*
EN:	FLOWER SHOW JIGSAW PUZZLE Heye Puzzles 1000 Pieces Bright Lily (NEW FOR 2016)
DE:	Carletto 3329739 Bright Lily Standard 1000 Teile SV
trans.:	*Carletto 3329739 Bright Lily Standard 1000 Pieces SV*

latent semantic analysis (LSA for short, also known as LSI – latent semantic indexing) has been applied on concatenated multi-lingual documents for cross-lingual document IR. Vinokourov et al. [27] uses canonical correlation analysis (CCA) for cross-lingual semantic text representation.

In text classification word representations that map words (or tokens) into vectors in a common vector space are commonly used. The works of Mikolov et al. [11], Pennington et al. [16], Turian et al. [25] and Peters et al. [17] have delivered high-quality word representations – word embeddings – induced through neural networks trained on monolingual data. Word embeddings have proven to be effective in numerous NLP tasks, such as sentiment analysis, textual entailment, and MT. Lai et al. [8] exploit pre-trained Skip-gram word embeddings for their Recurrent Convolutional NN (CNN) models for text classification. Word embeddings have also been successfully used for twitter sentiment classification [21, 24].

Mueller and Thyagarajan [13] present an RNN adaptation of a Siamese architecture [2, 7] with Long Short-Term Memory (LSTM) units [6] for computing text similarity. Another LSTM approach for the task of text entailment is presented in [19]. To improve the performance of their network, the authors exploit a word-to-word attention mechanism. Attention mechanisms have been successfully used for NLP tasks such as machine translation [1, 10], sentence summarization [20] and digit classification [12].

In our work, we are driven by a real-world application scenario. Therefore, we aim at a system that is not only robust, and with high predictive capabilities, but is also optimized towards speed and code sustainability in a large commercial environment. We draw a road map over different LSTM RNN network methods with and without attention as well as with and without pre-trained embeddings. We experiment with original as well as with machine translated data.

3 USE-CASES

Measuring the similarity of entries (or items) in cross-language e-commerce catalogues is essential for aligning products in the catalogue trees. Identifying which items represent the same product(s) across catalogues in different languages is fundamental for three use-cases:

UC1 **find in-domain comparable data.** Identifying the same or comparable catalogue entries is a way to create parallel corpora for training domain-specific MT engines. Given the high volume of data being published on a daily basis on e-Commerce websites, such corpora could encapsulate enough parallel text for high-quality MT. In addition, the organization of products in hierarchical catalogues allows data to be categorized comparatively from domain-specific to more general-domain responding to different MT requirements [28].

UC2 **synchronize content**. Sub-parts of product descriptions across language sites can be used for complementary knowledge exchange (e.g. by using MT to fill or enhance missing parts) thus improving the quality of product descriptions.

UC3 **detect erroneous/misaligned entries**. Automatically detecting erroneous/misaligned entries would help e-Commerce vendors to further improve the cohesion of their catalogues.

4 DATA

We considered two catalogue trees by eBay, one in English (EN) and one in German (DE). The catalogue entries contain the title of the item for sale (or an *item title*) with a maximum length of 80 UTF-8 characters (which can be noisy with characters representing emojis, for example). Table 1 shows examples of item titles together with their human translations. Next to the title for the item itself, we have access to other meta-information provided by the seller such as colour, quantity, the manufacturer, or other product specifics. However, their precision and coverage over the whole data is not complete.

To gather parallel training, test and validation data, we used the 12-digit universal product code (UPC) as well as its superset European Article Number (EAN-13) (again, entered by the sellers) to extract aligned items. The UPC and EAN numbers are unique per product and are shared among catalogues in different languages.

For each item, we also know which category (cars, toys, books, etc.) they are placed in, and we know from preliminary experiments that not all of them offer a fair challenge: for example, movies have a strong localization bias (e.g. "Soylent Green" has the title "... Jahr 2022 ... die überleben wollen" – "... year 2022 ... those that want to survive" when backtranslated from German), whereas music CDs are often verbatim. Thus, we restrict ourselves to the categories *Home and Gardening*, *Toys* and *Cameras & Photos*. We present details about our data in Table 2.

Table 2: Summary of the training/test/validation data.

category	reason for selection	# unique UPCs	# vocabulary entries for EN	DE	EN + DE
Home and Gardening	large variety in the items / brands	800k	105,429	65,124	129,490
Toys	high amount of used items with more user-generated content	800k	99,862	61,586	121,351
Cameras & Photos	lots of technical product descriptions	160k	30,997	23,956	42,335

5 APPROACHES

First, we built three neural models in a language-agnostic way to compute the similarity between item titles and identify whether they are the same item in the different catalogues. Second, we used MT to reduce the cross-lingual problem to a monolingual one and then retrain the models. As stated in previous work, such a language-aware approach depends on the quality of the MT system [4, 23]. This MT system is used in the production environment of eBayfor the English-German direction and has been optimized on title content.

We also experimented with and without pre-trained embeddings. When this is not made explicit in the following of this paper, we assume the embeddings are trained from our parallel data.

5.1 Language-agnostic similarity

In order to identify aligned items without considering the language as a factor, i.e. language-agnostic, we implemented three neural approaches:

ClassifierCat. We concatenate two sequences to form a joint input sequence that is given to a bidirectional LSTM RNN, in which the last hidden state is used in a soft-max layer for classification. The network predicts a probability distribution over n classes; the highest probability indicates to which of the n classes the input belongs. For UC1 and UC3, $n = 2$, same/different, while for UC2, (text synchronisation based on text entailment task), $n = 3$, e.g. positive entailment / negative entailment / contradiction. Our implementation is generic enough to allow both these tasks to be handled.

Siamese. A Siamese neural network combines two (or more) networks that have the same architecture and share the same weights, each of which takes as input one of two (or more) input sequences independently. It has already been successfully applied for text similarity in [13, 14, 26].

At training time the network parameters are optimized to compute a similarity score that would minimize the loss (in our case, mean squared error by default) – more similar input sequences will have a higher score. At test time, the output of the network is simply the similarity score between the input sequences. We use a distance metric, i.e. Euclidean distance, to compute this score. That is, in a multidimensional vector space, the Euclidean distance between the representations of the input sequences expresses their similarity (the smaller the distance, the higher the similarity).

Our Siamese architecture is focused on computing the similarity between two input sequences distance, i.e. at prediction time it will compute a value stating how similar the inputs are. While this is very suitable for UC1 and UC3, it is not suitable for UC2 as it will not handle a third dimension of comparison, as is the case of textual entailment. Accordingly, this approach is used only for handling UC1 and UC3.

ClassifierAttn. On long sequences LSTMs do not perform well as they need to compress all the information of a sequence in one context vector, i.e. the last state of the network. To solve this problem, attention mechanisms have been introduced [10] which allow the network to focus on parts of the sequence(s) that have the greatest importance. We implemented two attention mechanisms: (i) word-by-word attention inspired by [19] which we refer to as AttentionRTE[2], and (ii) Soft Dot Attention which we refer to as AttentionDot. We use soft attention instead of hard attention as we aim to provide a smooth representation of the encoded sequence where the important points are weighted accordingly, rather than select only a single point of interest and ignore the rest. Furthermore, we select dot attention as it is very fast (e.g. compared to additive attention mechanisms) and has shown to be very effective. The implementation of the ClassifierAttn model is similar to the ClassifierCat when it comes to the underlying LSTM network(s). However, there are two networks instead of one; two separate input sequences are provided: one from the L1 catalogue and the other from the L2 catalogue; also while for the ClassifierCat and the Siamese models a joint vocabulary is used, for the ClassifierAttn we use two different vocabularies: one for each language.

During the preprocessing step (prior to training the NN models), each sentence is tokenised and a beginning-of-sentence and end-of-sentence tokens (<bos> and <eos> respectively) were added to identify these positions. In addition, when joining two sentences for the ClassifierCat a <break> token was used inbetween to identify the joining point.

5.2 Language-aware similarity

The ClassifierCat approach and the Siamese approach use a shared embedding across both languages. We can thus compare their language-agnostic performance against a data set where the English titles are translated by a production facing English–German title translation service provided by eBay.

For the attention approach, the architecture contains separate embeddings for source and target. For this, we conducted early experiments using word2vec [11] embeddings trained on the individual languages, using MT on the tokens, and applying CCA to transform the English embedding space into the German. Then, we initialised the ClassifierAttn embedding layer with these pre-trained embeddings. This approach is similar to [3].

[2]We refer to this model as RTE, as the attention we use follows the work on Reasoning about Textual Entailment (RTE)

Table 3: Performance on the test set for all methods for all data (first four rows: combined; below: per category).

Category	Method	Time [min]	Epochs	Time/Epoch [min]	Precision	Recall	F1
all	ClassifierCat	440	20	22.0	0.955	0.974	0.964
	Siamese	243	13	18.7	0.837	0.961	0.896
	AttentionDot	455	17	26.8	0.962	0.944	0.953
	AttentionRTE	470	17	27.6	**0.962**	**0.972**	**0.967**
Toys	ClassifierCat	167	21	7.9	0.922	0.979	0.950
	Siamese	233	34	6.85	0.741	0.881	0.805
	AttentionDot	300	33	9.0	0.966	0.990	0.978
	AttentionRTE	314	34	9.2	0.955	0.981	0.968
'Cameras & Photos'	ClassifierCat	93	24	3.8	0.963	0.978	0.970
	Siamese	40	10	4.0	0.500	0.992	0.665
	AttentionDot	60	13	4.6	0.945	0.971	0.958
	AttentionRTE	73	16	4.5	0.958	0.980	0.969
'Home & Gardening'	ClassifierCat	139	28	4.9	0.887	0.952	0.918
	Siamese	155	33	4.7	0.882	0.92	0.903
	AttentionDot	170	29	5.9	0.958	0.964	0.960
	AttentionRTE	132	22	6.0	0.957	0.978	0.967

Table 4: Results on the combined three categories using additional language knowledge via MT and multi-lingual embeddings

Method	Precision	Recall	F1
ClassifierCat	0.955	0.974	0.964
+ machine translation	0.958	0.972	0.965
Siamese	0.837	0.961	0.896
+ machine translation	0.925	0.976	0.950
AttentionDot	0.962	0.944	0.953
+ multi-lingual embedding	0.912	0.970	0.940
AttentionRTE	**0.962**	**0.972**	**0.967**
+ multi-lingual embedding	0.927	0.949	0.938

5.3 Implementation details

Our implementation uses PyTorch[3] and gensim[4] as packages for neural network support and embeddings, respectively. The toolkit we developed within the scope of this work consists of a pipeline with the following component classes:

(1) data handling: to handle the large volume of data we implemented a set of scripts that (i) ingest the hadoop output; (ii) extract information per field from the hadoop output; (iii) align data based on defined field, e.g. UPC or EAN number; (iv) filter non-unique tuples; and (v) convert the data into a suitable format for each of the aforementioned methods.
(2) operational components: these include scripts for invoking training models with the aforementioned methods and testing with these models.
(3) pipelines: we developed several pipelines that would wrap different scripts together in order to provide an end-to-end functionality (i) data alignment (based on UPCs etc.); (ii) training word2vec, CCA, ClassifierCat, ClassifierAttn, Siamese; (iii) machine translation; (iv) training data preparation (including splitting into training, test and development sets) and (v) visualisation of results.

We also implemented a docker version of our toolkit; the quality of our implementation was continuously controlled through a series of regression tests.

[3]https://pytorch.org/
[4]https://radimrehurek.com/gensim/

6 EXPERIMENTS

From the data entries, we randomly assigned English entries with their German counterparts whenever they share the same UPC. For the experiments in this paper, we make the assumption that the UPCs are already known to the system. This means that development set and the test set contain entries from a withheld 2% of the parallel data (30K positive matches out of 145K for all categories) i.e. this pair match has not been encountered in the training. We created negative samples of double the size by randomly assigning titles from the same category but with a different product code.

For a fair comparison, we limited the embedding size of all methods to 100, and kept the number of hidden dimensions consistent to 50. Training was conducted with a batch size of 64, a patience of 5 and a maximum number of epochs of 100 (which was never exhausted in any training setting).

We summarize our experimental results for the language-agnostic case in Table 3; our results for the language-aware case are presented in Table 4.

Quality. First, our results for the language-agnostic case show that the attention-based models and especially the AttentionRTE outperform both the ClassifierCat as well as the Siamese network for the '**Toys**' and '**Home & Gardening**' categories as well as for **all** the data. The differences on the precision, recall and F1 metrics for the ClassifierCat, the AttentionDot and AttentionRTE models are quite small – for precision between 0.007 and 0.071; recall: 0.110 and 0.300 – indicating that these systems perform in a similar way. However, the observed difference between the AttentionDot,

Table 5: Attention for "3 packs 30 photos black frame fujifilm fuji instax mini film polaroid 7s sp 1"

attention	0.00	0.01	0.03	0.49	0.12	0.04	0.03	0.04	0.00	0.11	0.02	0.12
English token	<sos>	3x	fuji	instax	mini	black	frame	schwarzer	rahmen	sofortbildfilm	sofortbild	<eos>

Table 6: Excerpt attention for UPC-aligned "12 en 1 professional kit accessories bundle for gopro hd hero 4 3 2 1 sj4000"

attention	[..]	0.01	0.16	0.04	0.00	0.29	0.14	0.00	0.03	0.01	0.08	[..]
English token	[..]	kit	von	zubehör	für	gopro	hero	4.00	silber	schwarz	hero	[..]

AttentionRTE, the ClassifierCat and the Siamese networks range between precision: 0.076 and 0.463; recall: 0.021 to 0.109; and F1: 0.064 to 0.305, with the Siamese networks always underperforming. Furthermore, the Siamese network did not converge as well, when trained on the **'Cameras & Photo'** data. One reason for this might be that the similarity between items in a given category is too high for proper discrimination. Especially in the **Camera & Photo** category, where items have a distinct product code even for subtle changes in the title (obviously, "New Camera BRAND-XYZ Black model 5" and "New Camera BRAND-XYZ Black model 6" are always distinct product codes), the Siamese network fails to converge at all.

For the MT-supported experiments, while the impact on the ClassifierCat was negligible, our results show that MT aids the Siamese network. In particular the difference in scores for the Siamese network in comparison to the other methods is reduced for precision from 0.125 to 0.037, for F1 from 0.071 to 0.017 and in terms of recall, it performed best.

For the attention-based methods, we pre-computed word2vec skip-gram embeddings on the training parts, with the same size as the methods would use, i.e. 100. Then, we trained a linear CCA transformation via token-level MT system on the proportion where the MT could be linked to a target token. After applying the transformation on the whole embedding space, we then seeded the built-in embedding layer of the AttentionDot and the AttentionRTE with this shared embedding space. In our setting, this did not improve the performance (cf. Table 4), at least not in a setting where the pre-trained embedding is drawn from the same material as the actual classification.

Performance. On an Nvidia M40 GPU, our simplified models are quite quick to train as is clear from the measured time in Table 3. Even when using all the combined data, the maximum total training time is 470 minutes; maximum time per epoch is 27.6 minutes. It is also obvious that the attention-based methods are slower, but more robust. These times are very promising for a real-world application.

Note on attention. The attention-based methods have the benefit of yielding extra information on a word-level. While a detailed analysis of their performance is beyond the scope of this paper, we anecdotally found that attention puts more weight on product names than on brands; this behaviour is expected, as these are the trigger words that most often make the difference. We present two examples in Table 5 and Table 6. It is interesting to see that some German prepositions like "von" (of/from) or "mit" (with) gain high attention as well; we believe that the attention mechanism learns to identify these trigger words when associated with a product title. This seems beneficial to our case study since accessories to products have their own UPC and could be identified with these trigger words making it easier to differentiation between "a case *for* Samsung" and "a case *for* iPhone".

7 CONCLUSIONS AND FUTURE WORK

In this paper, we described and evaluated three neural approaches for cross-language item title alignment. The initial experiments are encouraging. We showed that the methods work fast and reasonably well on this data set.

As future work, apart from embarking on the typical journey of feature engineering, we intend to increase the data challenge by limiting the development and the test set to entirely unseen UPCs (currently, we only ensure that no title pair was encountered in training).[5] In addition, we aim to use informed negative sampling (measured by, e.g. catalogue tree approximity).

For the language pair itself, additional focus to the language specifics could be applied, such as compound splitting for German and normalization of abbreviations/units.

While our preliminary experiments with pre-trained cross-language embeddings did not yield overall better result than the strongest attention-based systems we plan to further investigate this topic by experimenting with different embeddings. This approach is not limited to bilingual data and could be applied to a much larger monolingual data collection before training a shared embedding space. Among others we consider investigating the applicability of MUSE[6].

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2015)*. San Diego, CA, USA. 15pp.

[2] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS 1993)*. Denver, Colorado, USA, 737–744.

[3] Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of theAssociation for Computational Linguistics, EACL 2014*. Gothenburg, Sweden, 462–471.

[5] Given the high scores we have seen on our test sets, UPC-distinction seems to be the most dire next step. Since the constraint on the UPCs leads to different data splits, the results are not comparable; however, as a first impression, early experiments indicate accuracy scores around 0.91 (ClassifierCat) and 0.93 for the Attention-based methods.

[6] https://github.com/facebookresearch/MUSE

[4] Bo Fu, Rob Brennan, and Declan O'Sullivan. 2009. Cross-Lingual Ontology Mapping – An Investigation of the Impact of Machine Translation. In *The Semantic Web*, Asunción Gómez-Pérez, Yong Yu, and Ying Ding (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.

[5] Jyoti Guha and Carmen Heger. 2014. Machine translation for global e-commerce on eBay. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, Vol. 2. Vancouver, Canada, 31–37.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. In *ICML Deep Learning Workshop*.

[8] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. [n. d.]. Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI 2015)*. Austin, Texas, USA, 2267–2273.

[9] Michael L Littman, Susan T Dumais, and Thomas K Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-language information retrieval*. Springer, 51–62.

[10] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2015)*. Lisbon, Portugal, 1412–1421.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[12] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS 2014)*. Montreal, Canada, 2204–2212.

[13] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*. Phoenix, Arizona, 2786–2792.

[14] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning Text Similarity with Siamese Recurrent Networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP, (Rep4NLP@ACL 2016)*. Berlin, Germany, 148–157.

[15] Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of Statistical Machine Translation Model for Cross-lingual Information Retrieval in a Service Context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*. Avignon, France, 109–119.

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2014), A meeting of SIGDAT, a Special Interest Group of the ACL*. Doha, Qatar, 1532–1543.

[17] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT 2018)*, Vol. 1. New Orleans, Louisiana, USA, 2227–2237.

[18] Natalia Prytkova, Gerhard Weikum, and Marc Spaniol. 2015. Aligning Multi-Cultural Knowledge Taxonomies by Combinatorial Optimization. In *Proceedings of the 24th International Conference on World Wide Web Companion, (WWW 2015) - Companion Volume*. Florence, Italy, 93–94.

[19] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2015. Reasoning about Entailment with Neural Attention. *CoRR* abs/1509.06664 (2015). arXiv:1509.06664 http://arxiv.org/abs/1509.06664

[20] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2015)*. Lisbon, Portugal, 379–389.

[21] Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado, 464–469.

[22] Steve Sloto, Ann Clifton, Greg Hanneman, Patrick Porter, Donna Gates, Almut Silja Hildebrand, and Anish Kumar. 2018. Leveraging Data Resources for Cross-Linguistic Information Retrieval Using Statistical Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, (AMTA 2018)*. Boston, MA, USA, 223–233.

[23] Dennis Spohr, Laura Hollink, and Philipp Cimiano. 2011. A Machine Learning Approach to Multilingual and Cross-Lingual Ontology Matching. In *The Semantic Web – ISWC 2011*, Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Noy, and Eva Blomqvist (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 665–680.

[24] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*. Baltimore, Maryland, USA, 1555–1565.

[25] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Uppsala, Sweden, 384–394.

[26] Nicola Ueffing, José GC de Souza, and Gregor Leusch. 2018. Quality Estimation for Automatically Generated Titles of eCommerce Browse Pages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT 2018)*. New Orleans, Louisiana, USA, 52–59.

[27] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. 2002. Inferring a Semantic Representation of Text via Cross-language Correlation Analysis. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS 2002)*. Vancouver, British Columbia, Canada, 1497–1504.

[28] Jian Zhang, Xiaofeng Wu, Andy Way, and Qun Liu. 2016. Fast Gated Neural Domain Adaptation: Language Model as a Case Study. In *Proceedings of the 26th International Conference on Computational Linguistics, (COLING 2016)*. Osaka, Japan, 1386–1397.

Product Phrase Extraction from e-Commerce Pages

Artem Vovk
avovk@google.com
Google
Mountain View, CA

Dmitrii Tochilkin*
dmitry.tochilkin@skolkovotech.ru
Skolkovo Institute of Science and Technology
Russia

Pradyumna Narayana
pradyn@google.com
Google
Mountain View, CA

Kazoo Sone
sone@google.com
Google
Mountain View, CA

Sugato Basu
sugato@google.com
Google
Mountain View, CA

ABSTRACT

Analyzing commercial pages to infer the products or services being offered by a web-based business is a task central to product search, product recommendation, ad placement and other e-commerce tasks. What makes this task challenging is that there are two types of e-commerce product pages. One is the single-product (SP) page where one product is featured primarily and users are able to buy that product or add to cart on the page. The other is the multi-product (MP) page, where users are presented with multiple (often 10-100) choices of products within a same category, often with thumbnail pictures and brief descriptions — users browse through the catalogue until they find a product they want to learn more about, and subsequently purchase the product of their choice on a corresponding SP page. In this paper, we take a two-step approach to identifying product phrases from commercial pages. First we classify whether a commercial web page is a SP or MP page. To that end, we introduce two different image recognition based models to differentiate between these two types of pages. If the page is determined to be SP, we identify the main product featured in that page. We compare the two types of image recognition models in terms of trade-offs between accuracy and latency, and empirically demonstrate the efficacy of our overall approach.

CCS CONCEPTS

• **Information systems → Computational advertising**; **Content match advertising**; **Information extraction**; *Structured text search.*

KEYWORDS

information extraction; web page classification; text classification; neural networks; natural language processing; computer vision

*This work was conducted while doing internship at Google.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316608

ACM Reference Format:
Artem Vovk, Dmitrii Tochilkin, Pradyumna Narayana, Kazoo Sone, and Sugato Basu. 2019. Product Phrase Extraction from e-Commerce Pages. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316608

1 INTRODUCTION

The total volume of e-commerce sales in the US is estimated to be around $500 billion annually today and it is still showing double-digit growth year-over-year [5]. In this growing e-commerce domain, Natural Language Processing (NLP) is playing an increasing role in different types of tasks. Some representative tasks include: (a) Product Matching, where the description of an item offered by the seller has to be automatically mapped to the appropriate product listing category — the NLP challenges here include handling heterogeneous ways of describing the same data (ranging from unstructured sparse text to rich structured text), erroneous information extraction leading to product duplicates, bundling multiple different products together in the same offer, etc., which have been recently tackled by deep neural network models [17]; (b) Automatic title generation for product pages, where titles on product browse pages have to be generated automatically — the main NLP challenge here is handling multiple languages, which has been tackled recently by sequence to sequence neural networks [13]; (c) Review summary and user profile generation — the core NLP task here is extracting representative sentences from product reviews and user profiles, for which recently neural networks have been used for aspect extraction and embedding [15]; and (d) Sentiment analysis from product review data, where the challenge lies in detecting the polarity of user sentiments in reviews by handling ambiguous text at different levels (e.g., sentence or paragraph level) — different machine learning models like Naive Bayes, Random Forest and Support Vector Machines have been applied on this task and compared [9]. In addition, on the advertising front, Pryzant et. al (2018) recently used different neural network modeling approaches (e.g., direct residualization, convolutional adversarial techniques) to identify words in search advertising text that can influence users' click behavior, while accounting for confounders [16].

One area that has not been well studied to the best of the authors' knowledge is deep understanding of the content generated by advertisers, such as identifying products and services offered in different types of e-commerce web pages. There are billions of

commercial web pages, offering a wide variety of products ranging from apparel goods to real estate, and billions of dollars are spent by consumers while interacting with these pages. As the market and the number of items offered on e-commerce platforms grows, it is becoming increasingly important to have a good understanding of the products being offered on a web page. In recent work, Zheng et al. [20] automatically extract product attributes from natural language text given a product name — they extract the relevant attributes from product profiles using bidirectional LSTM models for sequence tagging and CRF models for enforcing tagging consistency. A similar modeling approach using LSTM-CRF had been used earlier by Kozareva et al. [12] for extracting product brand name, model, etc. from text segments — in this case, however, the authors analyzed shopping queries for this task, not product pages.

In our work we distinguish between two kinds of e-commerce web pages: single-product pages (or sometimes called leaf pages) and multi-product pages (or sometimes called non-leaf pages). In single-product (SP) pages, the page's primary focus is on a single, specific product, typically with a large image of a product that is accompanied by product descriptions and user reviews, with an option to purchase the item by adding it to the shopping cart. In multi-product (MP) pages, the page's primary focus is to give users multiple choices of products that belong to the same category, so that users can browse through wide selection of products until they find an item that interests them, at which time the user can visit the corresponding SP page with more details of the chosen product and possibly make a purchase.

In SP pages, identification of the product phrase is relatively straightforward, although there are still challenges, e.g., similar products featured on the same page can easily confuse any product phrase identification algorithm. In MP pages, identification of the exact products being offered is less useful as the advertisers' primary intention usually is to show users a collection of similar products, so that users can make an informed decision.

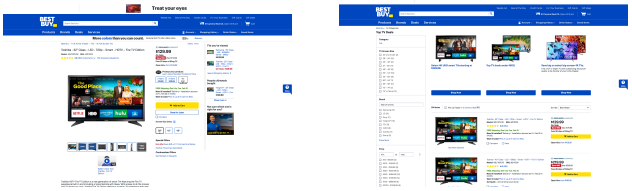
The classification of these two page types is of significant importance in e-commerce tasks, as it relates to how users can be matched to the commercial contents. For example, if the user has a clear intent of looking for a specific product, it makes sense for an internet search engine or a merchant to match them with a SP page for that product; whereas when the user is still in exploration or early phase of shopping, perhaps it makes more sense to match them with a MP page.

In this paper, we introduce three different approaches to classify product page type using different features and models. The following sections in the paper describe these approaches in detail, and outline the experiments we performed to analyze the trade-offs between various techniques.

2 APPROACH

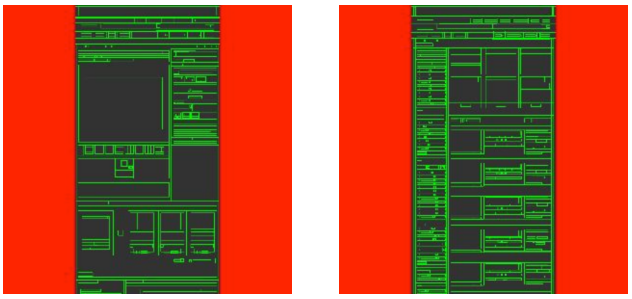
2.1 Page Type Classification

One approach of classifying page types is to render a web page as done by a browser to create an image of a page as seen by users. The goal here is to get an accurate representation of how users see web pages. However, the drawback is that the full rendering of each web page is quite resource intensive. In addition, it is not clear if the model is attributing the classification to certain products or other visual elements, rather than learning more abstract concepts. Figures 1(a) and 1(b) shows typical SP and MP pages to provide concrete examples.



(a) Single-product (SP) page (b) Multi-product (MP) page

Figure 1: Screenshot images. One product is prominently shown in the center of the page in SP page while each product is featured in one row of the block and users can scroll down to see more of similar products in MP page.



(a) Single-product (SP) page (b) Multi-product (MP) page

Figure 2: Wireframe images. A large block can be seen in the image in SP page while multiple similar blocks can be observed in MP page.

An alternate approach described in this paper is to generate lightweight web page structure representation. The idea is to depict the visual structure of the e-commerce pages without full scale rendering. The goal of this approach is to achieve the performance of image-based classification using high quality screenshots (as in our Web Render Service model), using a less resource-intensive page rendering mechanism that can make training and serving these models more efficient.

We first use internal tools to identify the render box of all blocks, which use different methods to identify the positions of the blocks on the screen [8, 10]. Then we use the OpenCV library [2] to render wireframe images of the render boxes. This allows us to create a simple wireframe representation of a web page that is unaffected by embedded images or other contents, and can classify the page purely by the page layout alone. Figures 2(a) and 2(b) show typical wireframe image of the SP and MP pages presented in Figures 1(a) and 1(b) respectively.

For both the image (screenshot) and wireframe representations, we used the Inception V2 model [19] trained on the JFT-300M dataset, which has more than 375M noisy labels (more than 19000 classes) for 300M images [7, 11, 18]. The last fully connected layer

weights of this model are randomly initialized and is fine-tuned using RMSProp optimizer, with learning rate of 5e-6, momentum (set to 0.9) and a random batch of size 48. The finetuning is stopped after 200,000 steps.

2.2 Main Product Phrase Classification

Once we know that the page has one main product (i.e., is a SP page), the next task is to identify what the product is. The purpose of the next model we will discuss is to classify phrases extracted from a web page as describing the product of the page or not. On a page that sells 'Nike Huarache Shoes', good product phrases could be for example 'Nike Huarache', while non-product phrases could include 'free delivery', 'heel support', 'Nike shoes review', etc. Since product/service names are usually noun phrases, only noun phrases extracted from the landing page are considered here. In this model, we use the candidate text (main product phrase), context sentence (a sentence containing the candidate text) and other metadata / features about the page to predict if the given phrase is a main product on the page. We also use the page title, topic of the page, url and salient terms found on the page as features. All text inputs are individually embedded using pre-trained word2vec embeddings (e.g., [14]). After that we used a CNN of width 2 to 5 each of which has 64 filters and 1-max pooling is performed over each feature map. On top of that we added a hidden layer of size 128 and before the last fully connected layer we fused all individual hidden layers into one layer by concatenating them.

2.3 Label/Dataset Generation

In this study, we make use of schema.org annotations [4] in both page type classification and main product phrase classification. According to the schema.org website, they provide "a collection of shared vocabularies webmasters can use to mark up their pages in ways that can be understood by the major search engines". Even though not all web pages have schema.org annotations, currently more than 1 million domain use Product annotations [3].

First, as we want to differentiate SP pages and MP pages, the most natural choice of the label is to count the number of products featured on the page and assign SP label if there is only one product and assign MP label if there are more than one, among all product pages. One way to achieve this is to use schema.org annotations. We collected some sample of search landing pages that contain https://schema.org/Product annotations — web pages that contain only 1 Product annotation are marked as a SP pages while the ones that contain more than 2 Product annotations are marked as MP pages. We discarded pages with exactly 2 Product annotations, as we have noticed from the eyeballing the data that there are a lot of mislabeled cases with 2 annotations. We also performed a human evaluation to estimate the quality of this labeling approach. We annotated 1k examples of each bucket. The results are summarized in Table 1 and they confirmed our concerns regarding pages with 2 Product annotations.

To avoid domain over-fitting, we limit to 5 pages per domain in our training data as well as we split our train/dev/test(80/10/10) sets by domain. In this way we were able to collect approximately 1.7 million labeled examples, 78% of which is SP and the remaining 22% is MP.

N	SP/MP	Label
1	93.0% / 4.3%	SP
2	45.3% / 51.4%	–
3+	31.4% / 61.0%	MP

Table 1: N is number of annotations and SP/MP shows human evaluated fraction of SP pages vs MP pages in each group. For example, among sample of web pages that has exactly 1 Product annotations, 93% of them are labeled as SP by human.

For the main product phrase classification, we use the phrase from the name property of product annotation as positive data. For the negative data, we use any other noun phrase found on the same web page. However, treating only the exact literal string match as the only truth is restrictive — it suffers lack of sufficient positive data as well as leads to many false negative results as well. In this study, we apply some rules to relax this. Let us suppose that the annotated phrase is "iconic silver charm **bracelet**" where the root of the phrase, bracelet, is shown in bold. The parse tree of this phrase looks like this, "[iconic] [[silver] charm] bracelet" showing "iconic" modifies "bracelet" as well as "silver" modifying "charm" and "charm" modifying "bracelet". Given this, any noun phrase sharing the same root as the annotated phrase, such as "PANDORA **bracelet**" or "iconic **bracelet**", would be considered positive. In cases where noun phrases contain an overlapping noun(s) but not sharing the same root, e.g., "silver charm" where "charm" is the root of the phrase, we discard these phrases as neither positive or negative examples. Finally, a noun phrase not sharing any noun or root token would be considered a negative example. This is demonstrated in Table 2. The dependency tree of each phrase is obtained using [6].

Phrase	Label
iconic silver charm **bracelet**	P
iconic **bracelet**	P'
silver charm **bracelet**	P'
PANDORA **bracelet**	P'
silver **charm**	I
iconic **store**	N

Table 2: Label of "P" is positive from the annotation (main product phrase), "P'" indicates induced synthetic positive (treated equally as true positive data) and "N" is negative. "I" is indefinite and discarded from training or evaluation. Root of each phrase is indicated by bold.

While generating training data we kept max 50 phrases per class/domain. We collected 85.7 million English phrases, 11.2% of which are product phrases and the rest 88.8% are non-product phrases.

3 RESULTS & DISCUSSIONS

First we evaluate our page type classification models using a sample of search landing pages in the held-out set.

Table 3 shows the precision, recall, F1 of the classifiers using threshold of 0.5 for the MP page class. Despite its simplicity, the wireframe image model presents comparable performance to the fully rendered image recognition model. The wireframe image model is multiple orders of magnitude faster than the full screenshot image model due to its simpler rendering, as seen in Table 3. We have observed memory consumption at the inference time is also significantly smaller in the wireframe model, offering superior characteristics for deploying in a production environment.

Approach	Precision	Recall	F1	Avg. Latency (ms)
Full Image	**0.89**	**0.88**	**0.89**	7,950
Wireframe Image	0.84	0.81	0.82	**230**[1]

Table 3: MP page classification using two different image recognition models. Wireframe approach, despite its simplicity, shows comparable metrics to full image rendering results at about 35 times faster processing time. Threshold of 0.5 is used.

Additionally we measured the quality of wireframe image classifier on 2k human labeled dataset (Table 4). We can see a significant drop in MP class precision, which can be explained by relatively low quality of our synthetically generated MP labels (as shown in Table 1) .However for our study we care more about the precision of SP classifier, which is relatively high (0.91).

Class	Precision	Recall	F1
Multi-product	0.69	0.83	0.75
Single-product	0.91	0.81	0.86

Table 4: Quality of classifier on 2k human labeled dataset. Threshold of 0.5 is used.

Next, we investigate how this page type classifier performs in each vertical. Table 5 shows the result of the classifier sliced by page vertical for the top 5 verticals that represent about 65% of the dataset. The vertical taxonomy and page type classification is taken from [1]. The model consistently perform well across different verticals. In particular the model has high precision (89% or higher) in identifying SP pages in most verticals. In fact, we see the average precision in bottom 5 categories is no different than the one in the bottom 5 categories (0.90), and out of 28 top level verticals we looked at, it shows 0.90 or higher precision on 22 verticals. This is useful for our current study as we are interested in identifying pages with a single product focus and be able to identify that product. Note that this evaluation is performed only on the pages that have the schema.org annotations.

Now armed with a precise selection of SP pages, we apply the main product phrase classification model to the web pages identified as SP pages. Table 6 shows the classification accuracy of the main product phrase classifier on a held out set.

We studied effects of different features used in the model by removing one or more features at the inference time. Table 7 shows

[1]This excludes pre-computation time of simplified layout rendering, since it can be easily cached in the production environment.

Vertical	P (MP)	R (MP)	P(SP)	R(SP)
/Shopping	0.87	0.85	0.92	0.93
/Home & Garden	0.83	0.78	0.90	0.93
/Autos & Vehicles	0.88	0.84	0.89	0.91
/Business & Industrial	0.79	0.73	0.89	0.91
/Computers & Electronics	0.81	0.75	0.89	0.92

Table 5: Slicing the web page type classification results by vertical. P indicates precision and R indicates recall, measure on respective class of MP or SP. Only the top 5 verticals from Google Cloud Content Categories [1] are shown, which represents approximately 65% of all eval data. All the results are from wireframe model.

Approach	Precision	Recall	F1
Product Phrase	.96	.94	.95
Non-Product Phrase	.94	.96	.95

Table 6: Main product phrase classifier results. Threshold of 0.5 is used.

this results. First, not surprisingly, it is clear that the phrase itself alone is not sufficient to differentiate the main product or not. There are some noun phrases that are clearly not that of a product. However there are many product phrases that are not the main product on the page, such as related items, which phrase itself is not sufficient to differentiate.

Ablated Feature	Loss in Precision
All but Phrase	-.23
Context	-.22
All but Phrase & Context	-.01
Page title	.00
Url	.00

Table 7: Feature ablation study at the inference time. Most of the predictive power comes from the phrase and its context sentence, and in fact only with these 2 features, it reaches to within 1 % of the best results.

Second, it is surprising that from other features only context sentence mattered for the classification. We took another look at our data and noticed that in a lot of cases context sentence for the product phrases is the product phrase itself. This is mostly due to the way we generated training data by using schema.org annotation: usually the annotated product phrase is a highlighted standalone sentence on the web page. This is not the effect we were looking for, since we also wanted to get the product phrases that located within sentences.

To overcome this issue we retrained our model without context sentence feature (Table 8). The quality of the new model is comparable to the previous one, we only lost 1 percent point in F1 product phrase score.

After that we did the ablation study again (Table 9). This time the most predictive features were the phrase itself as well as the

Approach	Precision	Recall	F1
Product Phrase	.95	.93	.94
Non-Product Phrase	.94	.96	.95

Table 8: Main product phrase classifier results without context sentence. Threshold of 0.5 is used.

title of the page. This result is aligned with our expectations, since the page title usually contains some highlights of the page, and for product pages the main highlight is the product itself.

Ablated Feature	Loss in Precision
All but Phrase	-.19
All but Phrase & Title	-.02

Table 9: Feature ablation study at the inference time of the model without context sentence. Most of the predictive power comes from the phrase and page title

4 CONCLUSION

In this paper, we presented two different image-based page structure detection models and a main product phrase classification model. The page type classifier model can be used to help us identify if the page focuses on a single product or multiple products. This allows us to extract a relevant product phrase from web pages where there is a single product focus. The main product phrase classifier can help us identify what that focused product is, among all the products mentioned on the page.

We would next like to combine the wireframe image recognition model with textual signals to understand the content better in the context of the presentation structure. In the future, we plan to combine this approach with information retrieval to detect and extract the most relevant information from a web page, with the help of web page structure.

ACKNOWLEDGMENTS

Authors would like to thank Ramakrishnan Srikant for thorough review of the manuscript.

REFERENCES

[1] 2018. Content Categories | Cloud Natural Language API | Google Cloud. https://cloud.google.com/natural-language/docs/categories
[2] 2019. OpenCV Library. https://opencv.org/
[3] 2019. "Product - schema.org". "http://schema.org/Product"
[4] 2019. schema.org. https://schema.org/
[5] 2019. U.S. Census Bureau News - QUARTERLY RETAIL E-COMMERCE SALES 3rd QUARTER 2018. https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf
[6] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally Normalized Transition-Based Neural Networks. In *Association for Computational Linguistics*. https://arxiv.org/abs/1603.06042
[7] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint* (2017), 1610–02357.
[8] Michael Xu Erik Hendriks and Kazushi Nagayama. 2014. Understanding web pages better. https://webmasters.googleblog.com/2014/05/understanding-web-pages-better.html
[9] Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data* 2, 1 (16 Jun 2015), 5. https://doi.org/10.1186/s40537-015-0015-2
[10] Google. 2018. Understand rendering on Google Search. https://developers.google.com/search/docs/guides/rendering
[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
[12] Zornitsa Kozareva, Qi Li, Ke Zhai, and Weiwei Guo. 2016. Recognizing Salient Entities in Shopping Queries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.
[13] Prashant Mathur, Nicola Ueffing, and Gregor Leusch. 2018. Multi-lingual neural title generation for e-Commerce browse pages. (2018).
[14] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. http://arxiv.org/abs/1301.3781
[15] Christopher Mitcheltree, Veronica Wharton, and Avneesh Saluja. 2018. Using Aspect Extraction Approaches to Generate Review Summaries and User Profiles. (2018).
[16] Reid Pryzant, Sugato Basu, and Kazoo Sone. 2018. Interpretable Neural Architectures for Attributing an Ad's Performance to its Writing Style. In *The 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 125–135.
[17] Kashif Shah, Selcuk Kopru, and Jean David Ruvini. 2018. Neural Network based Extreme Classification and Similarity Models for Product Matching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Association for Computational Linguistics. http://aclweb.org/anthology/N18-3002
[18] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *CoRR* abs/1707.02968 (2017). arXiv:1707.02968 http://arxiv.org/abs/1707.02968
[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*. http://arxiv.org/abs/1512.00567
[20] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. OpenTag: Open Attribute Value Extraction from Product Profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18)*.

Preliminary Investigation of Alleviating User Cold-Start Problem in E-commerce with Deep Cross-Domain Recommender System

Hanxin Wang
Osaka University
Suita, Osaka, Japan
wang.hanxin@ist.osaka-u.ac.jp

Daichi Amagata
Osaka University
Suita, Osaka, Japan
amagata.daichi@ist.osaka-u.ac.jp

Takuya Maekawa
Osaka University
Suita, Osaka, Japan
maekawa@ist.osaka-u.ac.jp

Takahiro Hara
Osaka University
Suita, Osaka, Japan
hara@ist.osaka-u.ac.jp

Hao Niu
KDDI Research, Inc
Chiyoda-ku, Tokyo, Japan
ha-niu@kddi-research.jp

Kei Yonekawa
KDDI Research, Inc
Chiyoda-ku, Tokyo, Japan
ke-yonekawa@kddi-research.jp

Mori Kurokawa
KDDI Research, Inc
Chiyoda-ku, Tokyo, Japan
mo-kurokawa@kddi-research.jp

ABSTRACT

Many current applications use recommender systems to predict user preferences, aiming at improving user experience and increasing the amount of sales and the usage time that users spent on the application. However, it is not an easy task to recommend items to new users accurately because of the user cold-start problem, which means that recommendation performance will degrade on users with little interaction, particularly for latent users who have never used the service before. In this work, we combine an online shopping domain with information from an ad platform, and then apply deep learning to build a cross-domain recommender system based on shared users of these two domains, to alleviate the user cold-start problem. Experimental results show the effectiveness of our deep cross-domain recommender system on handling user cold-start problem. By our framework, it is possible to recommend products to users of other domain through ad distribution in a more accurate level, and to increase sales amount of online shopping.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; **Recommender systems**; *Electronic commerce*; • **Computing methodologies** → **Neural networks**; *Learning from implicit feedback*;

KEYWORDS

Recommender Systems, Cross-domain Recommendation, Deep Learning, Implicit Feedback, E-commerce

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316596

ACM Reference Format:
Hanxin Wang, Daichi Amagata, Takuya Maekawa, Takahiro Hara, Hao Niu, Kei Yonekawa, and Mori Kurokawa. 2019. Preliminary Investigation of Alleviating User Cold-Start Problem in E-commerce with Deep Cross-Domain Recommender System. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316596

1 INTRODUCTION

In the era of information explosion, it is hard for users of online shopping services to explore all the available products in a limited amount of time. Thus, users prefer to get suggestions of products they might like to buy directly from online vendors. Because of this, recommender systems have become an important part of the e-commerce field, and have also been widely adopted by many other online services, such as online news [15], video sharing websites [1] and music streaming platforms [23].

The key component of the personalized recommender systems, which aims to model users' preferences for items based on their past interactions (such as purchase history), is collaborative filtering (CF) [19]. The most popular CF technique is matrix factorization (MF) [11, 14], which projects users and items to a shared latent space by factorizing a user-item interaction matrix. Given the immense success of deep learning on computer vision and natural language processing, some recent works have also employed deep neural networks (DNNs) for collaborative filtering (e.g. neural collaborative filtering [4]), or used neural networks to model auxiliary information (e.g., Word2vec [13]). With the powerful ability to learn a high-order nonlinear function, deep neural networks are suitable for learning complex relationship between users and items. However, both traditional MF methods and deep-learning-based methods suffer from the user cold-start problem, making it difficult to perform accurate recommendations when a new user comes, in

particular when recommending items to users who have never used the service before.

It is important to note that individual users are increasingly accessing a variety of different online services. For example, a user may buy products on an online shopping site and also reads news on other websites. An effective solution for the user cold-start problem is to transfer the knowledge from relevant domains and build a cross-domain recommender system. Since online shopping sites are empirically combined with ad platforms to promote the products, it is not difficult to collect a large amount of browsing histories in the ad domain of users who have bought products on an online shopping site. In this paper, we will use the term *bridge users* to refer to users who have both the purchase history in an online shopping domain and the browsing history in an ad domain.

Motivated by the above observation, we propose a framework to build the deep cross-domain recommender system by introducing side information in other domain. Specifically, we use a method to represent users by their browsing histories on an ad platform, and learn the user-item relationship in an online shopping service, based on the bridge users' purchase histories, to improve the recommendation performance for new users of the online shopping service, who have browsing histories on the ad platform. Through ad distribution, we can even accurately recommend products on online shopping sites to latent users who have never used the service before. We apply different kinds of CF models and perform extensive experiments on a real-world dataset, to show the effectiveness of integrating deep learning and side information from other domain for alleviating the user cold-start problem.

The rest of the paper is organized as follows. We briefly review related work in Section 2. Then we describe the basic information of the real-world dataset we used and our proposed framework in Section 3. The experimental setup and results are presented in Section 4. We finally conclude the paper and suggest some future work in Section 5.

2 RELATED WORK

There has been extensive study on applying deep learning to recommender systems. In this section, we review a set of representative studies related to our work.

The pioneer study on recommender systems, which used neural networks, proposed a two-layer Restricted Boltzmann Machine (RBM) to model users' explicit ratings on items [17]. Because of the strong performance of neural networks, deep learning has become a popular choice for building collaborative recommender systems [20, 21]. The user-based AutoRec [20] aims at learning a complex hidden structure to reconstruct a user's rating, given his or her historical ratings of other items. In order to avoid learning an identity function that always returns the same value as its argument, which performs very poorly on unseen data by general autoencoders, denoising autoencoders (DAEs) [21] have been used to learn from corrupted data, which involves artificially adding some noise to original input data, in order to train a more powerful model that can generalize to unseen data. Although these methods have shown the effectiveness of neural networks for CF, they focused on explicit ratings only, which means that they could fail to learn users' preferences based on positive-only implicit data, such as purchase history and movie viewing activity, because of the lack of negative feedback.

With respect to implicit feedback, there are also some recent works that applied deep learning to improve recommendation performance [4, 26]. [26] proposed a collaborative denoising autoencoder (CDAE) for CF with implicit feedback, which adds an additional user node to the input of the model compared with DAEs. [4] presented a state-of-the-art model called neural matrix factorization (NeuMF), which combines a generalized matrix factorization (GMF) model with a multi-layer feed-forward neural network, to model user-item interactions. However, CF is generally unable to handle new users and new items, which is known as the cold-start problem.

Solutions to the cold-start problem in recommender systems have mainly focused on new items (items which have no interactions with any users). This kind of works have primarily used DNNs to model auxiliary information, such as acoustic features of musics [23], images of items [28], and textual information of items [24]. In contrast to most traditional recommender systems, which are built within a single domain, several recent works have focused on combining data from multiple domains, which is called the cross-domain recommender systems. The approach is to utilize information of a shared set of users between different domains. For example, [6] proposed a method which improves the performance of apps recommendation by transferring knowledge from relevant news reading domain. However, this approach only improves the recommendation performance on shared users, and it is unable to deal with new users. In order to alleviate the user cold-start problem, [25] introduced user-user connections in social networking services (SNSs) to learn the embeddings of new users. We see that it is impractical to reconstruct the social networks for all the bridge users in online shopping site.

Our work combines the advantages of utilizing deep learning as CF technique and applying DNNs to extract users' and items' features. We use information from an ad platform domain, to build a cross-domain recommender system which alleviates the user cold-start problem in a real-world e-commerce site. More specifically, our framework performs more accurate recommendation for new users without previous interactions with the service.

3 FRAMEWORK

In this section, we first briefly describe our overall framework. We represent users by their browsing histories in an ad platform, with no user information in the online shopping domain, and represent items by their titles, subtitles, and descriptions through Word2vec. Then we train NeuMF based on the bridge users of these two domains, and do recommendation for users who have no interaction with the online shopping domain but with sufficient browsing records in the ad platform through ads distribution. Figure 1 illustrates the methodological procedure.

Next, we provide some basic information about the real-world dataset we used, and describe existing solutions for handling implicit feedback. Then, we outline the method to represent users and items in a latent low-dimensional space. Finally, we detail the NeuMF model that we use, and how we modify the model to fit our problem.

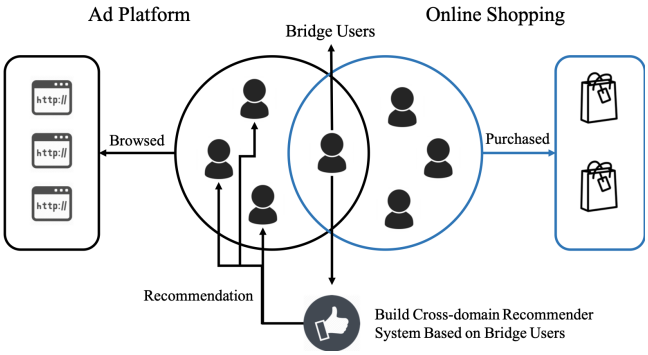


Figure 1: Methodological procedure of our framework

3.1 Learning from implicit feedback

We use a real-world online shopping dataset, an ad platform dataset, and a shared users dataset between these two domains.[1] The online shopping dataset contains users' purchase records from May 11th, 2017 to September 30th, 2017, and product information such as title, textual description, category, and price. In the ad platform dataset, there is a large amount of user browsing history over the same period as that of the online shopping dataset, and when a user accesses a web page where the ads are distributed, an access record will be stored in Ad platform. The shared users dataset is an ID mapping table of bridge users between the online shopping domain and the ad platform domain.

We build the cross-domain recommender system based on users' purchase records, which involves a kind of implicit feedback that can be defined as

$$y_{ui} = \begin{cases} 1, & \text{if user } u \text{ purchased item } i; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

It is important to note that a value of 1 for y_{ui} does not mean that user u actually likes item i, but only indicates that there has been an interaction between u and i, or u has interest in i. Similarly, $y_{ui} = 0$ does not necessarily mean that user u does not want to purchase this item, perhaps the user simply did not find the item before, which can be viewed as missing data. Given this fact, implicit feedback contains noisy signals about users' preferences, and naturally lacks negative feedback.

In addressing the problem of recommendation with implicit feedback, several previous works formulated the problem as predicting the scores of unobserved y_{ui}, which are used for ranking the items [5, 7, 16, 26]. Thus, the problem can be seen as learning $\hat{y_{ui}} = f(u, i|\theta)$ by optimizing an objective function, where θ denotes the parameters of the predictive model and f denotes an interaction function that represents the relationship between users and items. For the objective function, we follow [5, 7] to use point-wise loss, which aims at minimizing the point-wise loss between $\hat{y_{ui}}$ and y_{ui}, and sample negative instances from unobserved entries to deal with the absence of negative feedback [4, 5].

[1]Due to the policy of the data providers, we can not reveal the data sources.

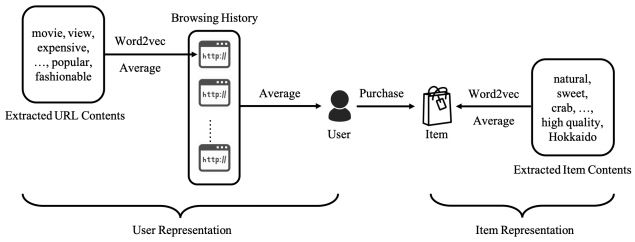


Figure 2: Process for user and item representation

3.2 Representation of Users and Items

Because our target is to recommend products in an online shopping domain to new users or people who have never used the service before, but with an abundant browsing history on an ad platform, we represent users only by corresponding browsing records in the ad platform but not information such as their profile in the online shopping domain, in order to generalize our recommender system to users in other domains. The process for creating representations of users and items is shown in Figure 2.

As different websites have different structures, and there are so many websites available, it is unrealistic to run crawlers for each website to collect contents, so we only extracted representative contents (title, keywords, and description) from web pages in the browsing records. In order to save time and reduce computational cost in the collection phase, we collect all the contents at the URL domain name level.[2] After obtaining a set of textual contents for each browsing record, we use Mecab[3] to do the Japanese text segmentation, and extract only nouns, verbs, and adjectives, filtering out meaningless words like stop words. Then we use a pre-trained Word2vec model to convert each word into a corresponding vector, and represent each browsing record as the average of these vectors. For each user, we similarly average the vectors of his/her browsing records as his/her representation.

We adopt a similar process for item representation, convert textual features such as title, subtitle, and description of each item into corresponding vectors using Word2vec.

3.3 Generalization of NeuMF

The neural matrix factorization model (NeuMF) is a state-of-the-art neural CF model that ensembles linear generalized matrix factorization (GMF) and non-linear multi-layer perceptron (MLP). Here, we first briefly introduce these two important components of the NeuMF, and then present a generalized version of NeuMF.

GMF. The GMF model is generalized from MF, which is the most popular model in the recommendation field. In general, MF takes the one-hot encoding of user ID and item ID as input, learns the latent embedding vectors, and finally calculates the inner product of the user's and item's latent vectors as the predicted score. The interaction function of MF can be defined as

[2]There is a problem with this method. It will fail to collect some important information that can distinguish different users more accurately. Although solving this problem might further improve the performance, we leave it as a future work.

[3]http://taku910.github.io/mecab/

$$y_{ui} = a_{out}(\mathbf{h}^T(\mathbf{p}_u \odot \mathbf{q}_i)), \tag{2}$$

where $\mathbf{p}_u$ denotes the user latent vector, $\mathbf{q}_i$ denotes the item latent vector, and $\odot$ denotes the element-wise product of the vectors. Also, a_{out} and $\mathbf{h}$ denote an identity function and a uniform vector with all elements as 1, respectively. If we generalize a_{out} to other activation functions such as sigmoid, and set $\mathbf{h}$ as weights that can be learnt from data, the MF model can be easily generalized to the more powerful and expressive GMF model.

MLP. The MLP is a multi-layer feed-forward neural network which is used to learn interactions between users and items. It is a more flexible model with a higher level of non-linearity than the GMF, which can only model user-item interactions linearly from the inner-product of two vectors. The MLP model under [4] is defined as

$$\begin{aligned} \mathbf{z}_1 &= \begin{bmatrix} \mathbf{p}_u \\ \mathbf{q}_i \end{bmatrix}, \\ \mathbf{z}_2 &= a_2(\mathbf{W}_2^T \mathbf{z}_1 + \mathbf{b}_2), \\ &\cdots\cdots \\ \mathbf{z}_L &= a_L(\mathbf{W}_L^T \mathbf{z}_{L-1} + \mathbf{b}_L), \\ \hat{y}_{ui} &= \sigma(\mathbf{h}^T \cdot \mathbf{z}_L) \end{aligned} \tag{3}$$

where $\mathbf{p}_u$ denotes the user latent vector, and $\mathbf{q}_i$ denotes the item latent vector, as in GMF. Furthermore, $\mathbf{W}_x$, $\mathbf{b}_x$ and $\mathbf{a}_x$ denote the weight matrix, the bias vector and the activation function for the x-th layer of MLP, respectively. For the $\mathbf{a}_x$, there are several available choices, such as sigmoid, tanh, and ReLU. As shown in [2], the sigmoid function can easily suffer saturation, which results in the vanishing gradient problem and halts the learning of the network. The tanh function alleviates the problems of the sigmoid function to some extent [12]. The ReLU has been proven to be non-saturated, well-suited for sparse data, and is capable of a higher convergence speed to speed up training. It can also reduce the overfitting problem of models [3]. Thus, we select ReLU as the activation function in our work, similar to several other works [4, 6, 27].

NeuMF. Figure 3 shows the structure of generalized NeuMF in our work. NeuMF combines GMF and MLP in their last hidden layers, for better modeling of user-item interactions by enabling them to mutually reinforce each other. As in [4], we allow GMF and MLP to learn separate embeddings under NeuMF, which is more flexible and expressive than sharing the same embedding between these two components. We adopt this setting and generalize NeuMF to fit our problem by using the representation of users and items described in Section 3.2 as input, in contrast to [4], which only uses the one-hot encoding of user (item) ID as input. The interaction function of generalized NeuMF is defined as follows:

$$\begin{aligned} \mathbf{z}_{GMF} &= \mathbf{p}_u^G \odot \mathbf{q}_i^G, \\ \mathbf{z}_{MLP} &= a_L(\mathbf{W}_L^T(a_{L-1}(\ldots a_2(\mathbf{W}_2^T \begin{bmatrix} \mathbf{p}_u^M \\ \mathbf{q}_i^M \end{bmatrix} + \mathbf{b}_2)\ldots)) + \mathbf{b}_L), \\ \hat{y}_{ui} &= \sigma(\mathbf{h}^T \cdot \begin{bmatrix} \mathbf{z}_{GMF} \\ \mathbf{z}_{MLP} \end{bmatrix}), \end{aligned} \tag{4}$$

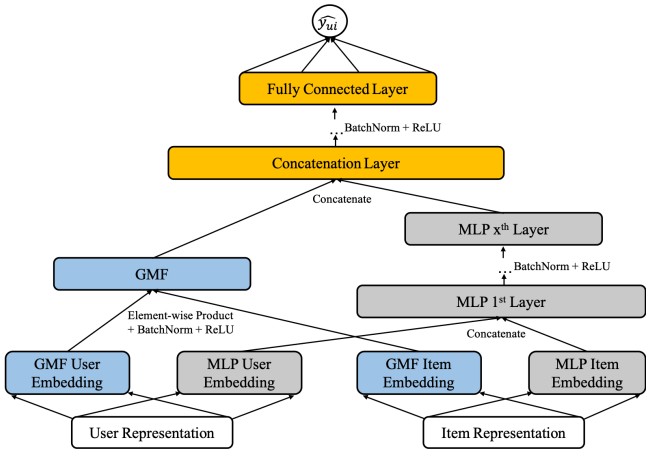


Figure 3: Structure of generalized NeuMF

where $\mathbf{p}_u^G$ and $\mathbf{p}_u^M$ denote the user embedding for GMF and MLP, respectively. Similarly, $\mathbf{q}_i^G$ and $\mathbf{q}_i^M$ denote the item embedding for these two components, and $\hat{y}_{ui}$ is the predicted result of NeuMF. We add a fully-connected layer after the concatenation layer to give NeuMF the ability to learn more complex user-item interactions. Then we apply batch normalization for each layer, in order to overcome the hard-to-converge problem, which might still be caused by ReLU, and speed up training [8].

For all the models described above, the objective function is defined as

$$L = -\sum_{(u,i)\in\mathcal{Y}\cup\mathcal{Y}^-} y_{ui}\log\hat{y}_{ui} + (1-y_{ui})\log(1-\hat{y}_{ui}), \tag{5}$$

where $\mathcal{Y}$ denotes the set of observed interactions and $\mathcal{Y}^-$ denotes the set of negative samples. This objective function is known as *logloss*.

4 EXPERIMENT

In this section, we conduct experiments to evaluate the performance of NeuMF in alleviating the user cold-start problem. Then we compare the NeuMF model with several baseline recommendation algorithms, to demonstrate the effectiveness of deep cross-domain recommender systems.

We first describe the setting of the real-world dataset that we used for training and evaluation. Next, we present the evaluation criteria, briefly review the baseline methods and the implementation details of all the methods that have been used in the experiments. Finally, we summarize the recommendation performance of the different methods, and discuss the results.

4.1 Experimental Settings

Dataset. As noted in Section 3.1, we used an ad platform dataset with users' browsing histories, and an online shopping dataset with users' purchase records from May 11th, 2017 to September 30th, 2017. We divide the purchase records into three parts: May 11th, 2017 to September 10th, 2017 for training, September 11th, 2017

Table 1: Statistics of the training set

Dataset	#interactions	#users	#items	Sparsity(%)
Training	156,287	14,659	18,511	99.942

Table 2: Statistics of the validation set and test set

Dataset	#interactions	#users	#items
Validation	12,410	9,937	1,539
Test	14,443	10,273	2,369

to September 17th, 2017 for validation, September 18th, 2017 to September 30th, 2017 for testing, and only use browsing histories from May 11th, 2017 to September 10th, 2017 for user modeling to avoid using future information to predict the past. As the original data is highly sparse, we only select users with at least five interactions (purchases) as a training set. The statistics are summarized in Table 1. In order to evaluate the recommendation performance for only new users, we filtered all the purchase records of users in the training set out of the validation and test sets. The statistics of these two datasets are shown in Table 2.

Evaluation Criteria. To evaluate the item recommendation performance, we follow the common strategy of top-K recommendation in [4, 6], to rank the target item of each interaction relative to items that the user did not interact with. Given that the items in our online shopping domain are only presented for a limited period of time, we determine that there are on average, about 1,500 items for sale each day from May 11th, 2017 to September 10th, 2017. In light of this, we randomly sample 1,499 items having no interaction with the user, and rank the 1,500 items, in contrast to [4, 6] which only sampled 99 items with no interaction for ranking. The performance of a ranked list can be evaluated by the *Hit Ratio* (HR) and *Normalized Discounted Cumulative Gain* (NDCG) [9]. The HR intuitively measures whether the target item is presented in the top-K list, while the NDCG accounts for the hit position by assigning higher scores to hits in the top ranks. As the online shopping site in our work always provide a list of recommended items with length between 5 to 10, we set K as integers from 5 to 10 for evaluation, calculate both metrics for each interaction, and present the average score.

Baselines. We compare the NeuMF employed in our work with various different methods:

- **Cosine Similarity.** This is the standard method to model user-item relationship, and is widely adopted as the final step in recommendation models.

- **ItemPop.** This is a method which ranks items based on their popularity. In our case, we judge the popularity by the sales amount of products that are still in sales. It is a non-personalized method which can be applied to cold-start users.

- **GMF [4].** As noted in Section 3.3, GMF is a generalized version of traditional MF. It is a shallow neural network model that learns user-item interactions as applying activation function to the linear combination of the element-wise product of the input vectors, which is more expressive than traditional MF.

- **DMF [27].** DMF is a state-of-the-art deep matrix factorization model that projects users and items into a low-dimensional latent space by DNN, and calculates the cosine similarity of these two vectors as the predicted result. We follow the result in [27] to build a DMF with two layers and tune the other hyper-parameters to produce the best performance.

Implementation Details. All the neural network models are implemented based on PyTorch[4]. For all the models except Cosine Similarity and ItemPop, we randomly sampled four negative samples for each positive instance. The parameters to be trained are randomly initialized from a Gaussian distribution $\mathcal{N}(0, 0.01^2)$. The optimizer is SGD with the batch size of 256, and we tested the learning rate of [0.001, 0.005, 0.01]. We also tested the number of neurons in the last hidden layer of [16,32,64,100]. For the structure of MLP model, we used a tower pattern, which halves the layer size for each successive higher layer. To optimize the hyper-parameters for each model, we tuned them based on the validation set.

4.2 Experimental Results

In this section, we report the recommendation performance of the various methods, and discuss the experimental results. Table 3 shows the performance of HR@K with K varied from 5 to 10. Table 4 shows the performance of NDCG@K in the same situation. Due to the weak performance of Cosine Similarity, it is omitted to better highlight the performance of other methods.

According to the experimental results, the shallow GMF model outperforms ItemPop in all the situations except HR@6 and HR@7, particularly for the great improvement in NDCG. We can see that deep-learning-based models, NeuMF and DMF, consistently give better performance in both metrics than the other methods, for all the values of K, with improvement up to 15.267% and 18.023% in HR@10 compared to the best shallow GMF model and ItemPop respectively. For NDCG@10, the performance of the best deep-learning-based model NeuMF is about 2.08 times better than the GMF and 3.64 times better than ItemPop. This shows the effectiveness of neural approaches in alleviating the user cold-start problem using knowledge introduced from another domain.

We can also find that NeuMF outperforms DMF and achieves 4.348% improvement in HR@10 and 1.04 times better than DMF in NDCG@10. It shows NeuMF is better at modeling the overall user-item interactions and can rank more target items into the top-10 ranked list than DMF, which is more practical because it is general to recommend a list of items to users. The results of NDCG@K also show this fact, as DMF still gives better performance than NeuMF in NDCG@5 and NDCG@6, while it has already underperformed NeuMF in HR@5 and HR@6.

5 CONCLUSION AND FUTURE WORK

In this work, we transferred information from an ad platform to an online shopping domain, based on bridge users of these two domains. Then we applied different kinds of CF models to build

[4]https://pytorch.org

Table 3: Performance of HR@K

K	5	6	7	8	9	10
NeuMF	**13.771%**	**17.725%**	**20.951%**	**23.167%**	**24.815%**	**26.165%**
DMF	12.394%	14.706%	16.769%	18.694%	20.425%	21.817%
GMF	6.003%	7.125%	8.032%	9.036%	10.005%	10.898%
ItemPop	1.461%	8.142%	8.142%	8.142%	8.142%	8.142%

Table 4: Performance of NDCG@K

K	5	6	7	8	9	10
NeuMF	0.0666	0.0807	**0.0914**	**0.0984**	**0.1034**	**0.1073**
DMF	**0.0728**	**0.081**	0.0879	0.094	0.0992	0.1032
GMF	0.0359	0.0399	0.0429	0.046	0.049	0.0515
ItemPop	0.0057	0.0295	0.0295	0.0295	0.0295	0.0295

cross-domain recommender systems, and evaluated their recommendation performance with respect to the new users of the online shopping domain. The experimental results show that deep-learning-based models far outperform other shallow models for CF, in alleviating the user cold-start problem. Moreover, NeuMF could model overall user-item interactions better than DMF in HR@K when K is between 5 and 10, which shows that the model combines linearity and non-linearity is more powerful than the pure non-linear model in the performance of our task.

In future, we will study different kinds of user-modeling methods, such as applying TF-IDF [18] to give different weights to different URLs and words, or directly converting URLs into vectors to represent users [22], in order to further improve the performance of our framework. In addition, we plan to further explore the impact of time on the recommendation performance, and investigate performance of these methods with respect to new items in the online shopping domain. Moreover, we are particularly interested in building cross-domain recommender systems through transfer learning methods like domain adaptation [10].

ACKNOWLEDGMENTS

This research is partially supported by JST CREST Grant Number J181401085.

REFERENCES

[1] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *RecSys*. 191–198.
[2] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*. 249–256.
[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *AISTATS*. 315–323.
[4] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.
[5] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR*. 549–558.
[6] Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. CoNet: Collaborative Cross Networks for Cross-Domain Recommendation. In *CIKM*. 667–676.
[7] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *ICDM*. 263–272.
[8] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167* (2015).
[9] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *TOIS* 20, 4 (2002), 422–446.
[10] Heishiro Kanagawa, Hayato Kobayashi, Nobuyuki Shimizu, Yukihiro Tagami, and Taiji Suzuki. 2018. Cross-domain Recommendation via Deep Domain Adaptation. *arXiv:1803.03018* (2018).
[11] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.
[12] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.
[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.
[14] Andriy Mnih and Ruslan R Salakhutdinov. 2008. Probabilistic matrix factorization. In *NIPS*. 1257–1264.
[15] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*. 1933–1942.
[16] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*. 452–461.
[17] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. 2007. Restricted Boltzmann machines for collaborative filtering. In *ICML*. 791–798.
[18] Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval. (1986).
[19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. 285–295.
[20] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *WWW*. 111–112.
[21] Florian Strub and Jeremie Mary. 2015. Collaborative filtering with stacked denoising autoencoders and sparse inputs. In *NIPS workshop on machine learning for eCommerce*.
[22] Yukihiro Tagami, Hayato Kobayashi, Shingo Ono, and Akira Tajima. 2018. Representation Learning for Users' Web Browsing Sequences. *IEICE Transactions on Information and Systems* 101, 7 (2018), 1870–1879.
[23] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *NIPS*. 2643–2651.
[24] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *KDD*. 1235–1244.
[25] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item silk road: Recommending items from information domains to social users. In *SIGIR*. ACM, 185–194.
[26] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *WSDM*. 153–162.
[27] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep Matrix Factorization Models for Recommender Systems. In *IJCAI*. 3203–3209.
[28] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *WWW*. 649–658.

Reference Product Search

Chu Wang*
Amazon.com
chuwang@amazon.com

Lei Tang†
Amazon.com
leitang@amazon.com

Shujun Bian
Amazon.com
sjbian@amazon.com

Da Zhang
Amazon.com
dazh@amazon.com

Zuohua Zhang
Amazon.com
zhzhang@amazon.com

Yongning Wu
Amazon.com
yongning@amazon.com

ABSTRACT

For a product of interest, we propose a search method to surface a set of reference products. The reference products can be used as candidates to support downstream modeling tasks and business applications. The search method consists of product representation learning and fingerprint-type vector searching. The product catalog information is transformed into a high-quality embedding of low dimensions via a novel attention auto-encoder neural network, and the embedding is further coupled with a binary encoding vector for fast retrieval. We conduct extensive experiments to evaluate the proposed method, and compare it with peer services to demonstrate its advantage in terms of search return rate and precision.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Search methodologies**; **Machine learning**.

KEYWORDS

Product Search, Representation Learning, Semantic Hashing, Attention Mechanism, Denoising Auto-Encoder

ACM Reference Format:
Chu Wang, Lei Tang, Shujun Bian, Da Zhang, Zuohua Zhang, and Yongning Wu. 2019. Reference Product Search. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316598

1 INTRODUCTION

Product modeling and retrieval tasks, including product advertising, substitution, and recommendation, are fundamental for e-commerce business. To help customer discover more products and improve shopping experience, e-commerce platforms like Amazon, eBay, Taobao, and JD all gradually launched new browsing or assistance features related to similar product comparison, alternative product recommendation, and substitute product suggestion. In addition to the traditional query-to-product search, such a product-to-product retrieval mechanism contributes significantly to the e-commerce business. For example, JD launched a "find similar" widget available for certain products, and Amazon also provides this widget for products in the browse history, to provide more alternatives closely related to a product of customers' interest. In general, the solutions consist of two stages: a candidate product set is retrieved first, followed by a task-specific ranking model to generate the results. Often, research interests focus on a ranking model built to optimize towards such a business application, but a suitable candidate product set is required to feed the ranking model and it is not well discussed in the literature.

Qualification of the candidate product set differs for different business objectives. In spite of those differences, the candidate product set is generally sourced from catalog information, behavioral data, and human annotations. The traditional inverted-index based retrieval relies on indices which are generated via product attribute tagging, and products sharing common indices like keywords are regarded as the candidates [12, 28]. If the products are assigned or classified in a taxonomy, the products under the same category can be used as the candidates, though mis-classified query product will lead to irrelevant results, and the size of the retrieved candidate set is uncontrollable and may be highly skewed. Customer behavioral information is another source to generate candidate products. Products that often viewed together or purchased together can serve as candidates for each other [19]. Though widely adopted in e-commerce business especially for product advertising and recommendation, the mentioned methods need to handle issues like feature sparsity, noisy data sources, low return rate, and inability to adjust the retrieval size. Ideally, the retrieval approach should be able to fetch a enough number of candidate products similar to the query, and the number of candidates should be flexible to adjust in order to accommodate various downstream ranking models for e-commerce applications.

In this paper, we focus on the problem of obtaining a set of reference products for a query product of interest. The reference products can be further fed to a ranking or relevance model to optimize the business objectives like clicks, conversions, or profits. We formulate this product retrieval problem as a search task, where we build product embedding vectors, quantify product similarity by vector distance, and conduct nearest neighbor search (NNS) in the vector space to surface the reference products. In addition to the requirement of flexibly adjusting retrieval size that NNS naturally provides, we focus on two metrics to evaluate the quality of the proposed search method. The precision of the search results measures

*† Both authors contributed equally.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316598

the quality of the top search results; the return rate estimates the ability to retrieve enough reference products for different queries.

There are various challenges involved to build a desired product-to-product search method. Customer behavior data is limited to only popular products, making it difficult to achieve high return rate. The catalog information is more widely available for product embedding, but feature extraction from low-quality or even missing product information is challenging [5, 36]. In fact, we observe 10% to 45% missing rates for several important product fields, and the available catalog information is plagued by poorly written catalog data with irrelevant or duplicated information. Large-scale vector search is also challenging in terms of the computation efficiency. Exact NNS is not realistic for moderately large datasets [33], and one need to refer to approximate methods in order to reduce the latency. For datasets with a large volume of items, the balance between the efficiency gain and the quality loss is difficult to handle.

In this paper, we propose a novel attention auto-encoder neural network to build high-quality and robust product vectors. In order to achieve better search precision and return rate, the negative impact of missing or low-quality attributes is minimized via attention mechanism [29]. The vector search is then conducted by an optimized semantic hashing algorithm [23], where each product embedding is coupled with a binary encoding so that NNS can be achieved by searching in the vicinity of the query encoding. The encoder is optimized to penalize tiny and huge binary buckets for better search precision and latency. Compared to existing product-to-product retrieval methods, the proposed method is able to obtain better precision while enhancing the return rate drastically. In our experiments for high-traffic products, the proposed method achieves 92.2% none-zero return rate compared to 86.5% from the top-performing existing method. This advantage is enlarged to 63.8% v.s. 3.9% for general products. As an approximate NNS method, our search method is able to achieve 90.7% recall rate against the exact 100-NNS compared to 74.2% or less recall from state-of-the-art packages. As for the computation efficiency, we achieve an average latency of less than 6ms for a product pool of 40 million products with a single machine of 61GB memory and a single Nvidia K80 GPU, and the method can scale up to support large product pools with multiple machines.

2 BACKGROUND AND OUR CONTRIBUTION

Product retrieval is fundamental for modeling and business applications in e-commerce. In order for relevance or ranking models to apply, various methods are adopted to retrieve relevant products, including the vastly used inverted-index based methods like Elasticsearch or Solr [12, 28]. In addition to supporting the downstream business applications, the retrieved products can also be consumed by instance-based product modeling [1, 35]. The retrieval method should be able to fetch enough products for a large proportion of the query products. The retrieval recall rate is also considered a good metric, but it is not commonly adopted because of the lack of ground truth [24]. In this paper, the proposed search method is largely related to two areas. Product representation learning and natural language processing help to convert a product to an embedding vector; the approximate nearest neighbor search supports similarity search based on product embeddings at scale.

Most of the existing product representation techniques for e-commerce depends largely on customer behavior data. Customer co-view and co-purchase information can be directly used to build product embedding such that products with similar browse or purchase history will share similar vectors [19]. Product catalog data, on the other hand, is a collection of free-form texts like title, description, brand, *etc.*. There are supervised and unsupervised ways to transform the catalog data into a vector space. The supervised way originates from ImageNet [8], where a multi-source neural network is trained to predict certain product categorization labels, and the last hidden layer is used as the instance embedding [5, 13, 27]. The categorization labels require extensive annotation while the embedding vector is found of high quality. The unsupervised way, on the other hand, applies to broader cases, especially when labels are unavailable. With the well established methods of word2vec and sentence2vec [7, 9, 16, 18], the problem of product representation learning effectively becomes the task of vector aggregation, namely the method of combining embedding vectors from multiple product fields. Though this problem seems fundamental for natural language processing in e-commerce, to the best of our knowledge, there is no related research dedicated to this problem.

Vector space search has drawn increasing attention because of its ubiquitous applications. In addition to product search, vector search is also frequently used in recommender systems [4, 17], and extreme classifications [30, 34]. Depending on the definition of the similarity metric, there are nearest neighbor search (NNS) on Euclidean distance, maximum cosine similarity search (MCSS), and maximum inner product search (MIPS) [26, 33]. While normalizing the embedding vectors will unify the three, there are independent works for each scenario [26]. The exact search method is not scalable since it requires calculating the pairwise similarity between the query and each of the candidates. Therefore, related works are focusing on approximate methods which can be roughly classified into tree-based approaches, hashing based approaches, and other approaches [3, 14, 15, 25, 26]. There are a large amount of works dedicated to high-performance approximate KNN, including FLANN, Iterative Expanding Hashing (IEH), Non-Metric Space Library (NMSLIB), and ANNOY [6, 11, 20, 21]. Our search method is a fingerprint-type method, belonging to the hashing based approach as IEH, while the other three methods are tree-based approaches.

We discuss technical details of the proposed search method in Section 3, followed by experiments regarding the return rate, the precision, and the recall according to exact KNN in Section 4. Section 5 discusses possible applications of the proposed method and concludes the paper. Before going into detailed discussions, we would like to highlight our contribution as follows:

- We have developed a product-to-product search method for reference product retrieval. The reference products are obtained via an optimized semantic hashing approach on product embedding generated by a novel attention auto-encoder neural network.
- With satisfactory precision, the proposed method is able to achieve considerably higher return rate compared to existing peers. Thus, our reference product search can support general downstream e-commerce applications.

- As an approximate KNN method, the proposed search algorithm is able to achieve high recall rate compared to state-of-the-art approximate KNN packages.
- It is flexible to adjust the number of returned candidate products based on the requirement of the applications. The latency-precision balance is adjustable in real-time to accommodate different use cases; the product embedding and retrieval encoding are also of a plug-and-play type.

3 METHOD

In this section, we first describe the offline and online processes of the proposed search method, followed by detailed discussions on product vectorizer and binary encoder that enable high-quality and fast search for products.

Two transformers are used to convert the text data of product catalog information into vector spaces: the product vectorizer $g(\cdot)$ and the binary encoder $h(\cdot)$. For a product p, the vectorizer converts the product catalog information into a d-dimension embedding (column-)vector $\vec{v} = g(p) \in \mathbb{R}^d$, and the binary encoder further transforms the embedding vector into a d'-dimension binary encoding vector $\vec{b} = h(\vec{v}) \in \mathbb{B}^{d'}$. We will introduce the design and training of the vectorizer $g(\cdot)$ and the encoder $h(\cdot)$ in Section 3.1 and 3.2, respectively. Let $\mathcal{P}$ denote the product pool where search results are retrieved from, and write $N := |\mathcal{P}|$ the size of $\mathcal{P}$. Vectorizing and the encoding processes are conducted offline for all the products in $\mathcal{P}$, resulting in an embedding set V and an encoding set B. For any binary encoding $\vec{b}_i \in B$, define the corresponding binary encoding bucket $\mathcal{B}_i := \{\vec{v} \in V \mid h(\vec{v}) = \vec{b}_i\}$. The set of all the buckets $\mathcal{B}^* := \{\mathcal{B}_1, \mathcal{B}_2, \dots\}$ defines a partition over V.

The online search process follows the semantic hashing mechanism [23]. We retrieve a small subset consisting of M ($M \ll N$) products from $\mathcal{P}$ via low-cost computation, followed by applying exact NNS to the candidate set to get the sorted search results. Such a schema approximates the exact NNS on the larger product pool $\mathcal{P}$ where M is adjusted to balance the quality of approximation and the latency. For any given query product p_q, our reference product search method proceeds as follows:

(1) Product vectorization. Calculate the query embedding $\vec{v}_q = g(p_q)$ and the encoding $\vec{b}_q = h(\vec{v}_q)$.
(2) Sort all binary buckets $\mathcal{B}_i$ in $\mathcal{B}^*$ according to the Hamming distance between its encoding $\vec{b}_i$ and the query encoding $\vec{b}_q$ to get a ranked list: $\mathcal{B}_{q_1}, \mathcal{B}_{q_2}, \mathcal{B}_{q_3}, \dots, \mathcal{B}_{q_{|\mathcal{B}^*|}}$.
(3) Find the minimal cut-off position c such that $|\mathcal{B}_{q_1}| + |\mathcal{B}_{q_2}| + \cdots + |\mathcal{B}_{q_c}| \geq M$.
(4) Conduct exact NNS for $\vec{v}_q$ and candidate set $\cup_{i=1}^{c} \mathcal{B}_{q_i}$ and return the top K ($K \leq M$) products.

Note that we conduct exact NNS for the retrieved M products and returns the top-K results as reference products. In reality, we usually apply a cut-off threshold γ to further filter out low quality products to ensure the precision of the search results. For a specific ranking or relevance model, the search results are further utilized or consumed. Intuitively, if the retrieved set covers a large proportion of the actual M-nearest neighbors of $\vec{v}_q$ in V, then the approximation should be good enough for downstream ranking models. We will discuss how to optimize the encoder $h(\cdot)$ for better search quality in Section 3.2.

Furthermore, we would like to highlight that the product pool can be expanded incrementally without retraining the vectorizer and the encoder. This is advantageous since the time-consuming offline pre-processing is not frequently conducted.

3.1 Product Representation Learning

The product vectorizer $g(\cdot)$ converts the product catalog data to a vector by incorporating signals from multiple attributes. A single product catalog field is essentially a sentence or a paragraph, and simple word2vec or sentence2vec model can be applied. Though there are supervised approaches to obtain product embeddings, we choose the unsupervised approach because it can be easily applied to billions of products without extensive human annotation efforts. We train a fastText [16] model on all the product information (around 400B tokens), and use this model as the field vectorizer. Each product catalog field is then converted into a vector $\vec{u} \in \mathbb{R}^d$.

The challenge is how to combine embedding vectors $\vec{u}^1, \vec{u}^2, \dots, \vec{u}^m$ from multiple fields. Taking more fields into consideration helps utilizing more signals from the product information, but naively concatenating field vectors suffers from two subsequent problems:

- The curse of dimensionality: more fields mean higher dimension of the product embedding, which increases latency.
- Missing fields and low-quality catalog information: we observe 10% to 45% missing rates for several important product fields, not to mention poorly-written fields with irrelevant or duplicated information.

Notice that the data quality issue does not come from data collection and processing. Therefore, it should be alleviated via proper modeling. To that end, we propose a novel attention auto-encoder network for embedding aggregation. The goal is that we smartly assign a weight $\alpha_j(U)$ to each field vector $\vec{u}^j \in \mathbb{R}^d$ so that the product embedding is a proper convex combination $\vec{v}_p = g(p) := \sum_{j=1}^{m} \alpha_j \vec{u}^j$, where the matrix $U := [\vec{u}^1, \vec{u}^2, \dots, \vec{u}^m]$. The averaging resolves the curse of dimensionality, and the embedding quality is optimized via minimizing information loss during the averaging. To be more specific, the weight vector $\vec{\alpha}(U)$ comes from a self-attention module defined as a softmax distribution:

$$\vec{\alpha}_j(U) := \exp(\phi(\vec{u}^j)) \Big/ \sum_{i=1}^{m} \exp(\phi(\vec{u}^i)), \tag{1}$$

where $\phi(\cdot)$ is a simple fully-connected neural network with one hidden layer and a scaler output: $\phi(\vec{u}^j) := \vec{\eta} \tanh(W\vec{u}^j + \vec{\xi})$. The attention module "looks at" each catalog field, returns a score estimating the relative amount of information from that field, and the softmax function further adjusts the weights for balance. The parameters $W, \vec{\xi}, \vec{\eta}$ are learned via minimizing the information loss from U to $\vec{v}$, or in other words, to ensure the combined vector $\vec{v}$ can best recover U from another two-layer neural network, defined as

$$\varphi(U) := \sigma\left(W^{(1)}\vec{v}_p + \vec{\xi}^{(1)}\right) = \sigma\left(W^{(1)}U\vec{\alpha}(U) + \vec{\xi}^{(1)}\right), \tag{2}$$

$$\vec{\hat{u}}^j(U) := W_j^{(2)}\varphi(U) + \vec{\xi}^{(2)}, \tag{3}$$

where the activation function $\sigma(\cdot)$ is the sigmoid function. The loss for the neural network is simply the mean squared error:

$$L := \frac{1}{m \times N} \sum_{p \in \mathcal{P}} \sum_{j=1}^{m} \left\| \vec{u}_p^j - \hat{u}_p^j(U) \right\|^2, \tag{4}$$

where the loss function L depends on matrices $W, W^{(1)}, W_j^{(2)} (1 \le j \le m)$ and vectors $\vec{\xi}, \vec{\xi}^{(1)}, \vec{\xi}^{(2)}, \vec{\eta}$. During training, U is generated from the catalog data of 40 million products, the dimension of the hidden layer is 32 for the attention module, 64 for $\varphi(U)$, and the model parameters are optimized via Adam Optimizer [24]. Note that for product embedding $g(\cdot)$, we only need the parameters $W, \vec{\xi}, \vec{\eta}$ in the attention module. The rest of the model parameters are auxiliary to help tune the attention module. For a given product with field embedding U, the attention module gives the optimal weights $\vec{\alpha}(U)$ so that the product embedding $\vec{v} = U\vec{\alpha}(U)$ can best incorporate and recover signals from U. The detailed neural network structure is demonstrated in Figure 1.

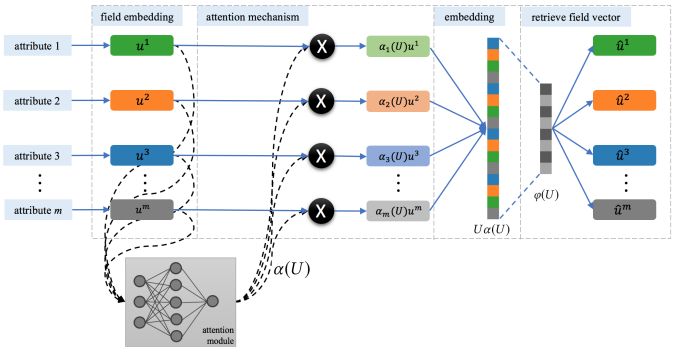


Figure 1: Model structure of the attention auto-encoder.

We conduct extensive experiments to evaluate the quality of embeddings based on the proposed attention auto-encoder (AAE) method. It is worth noting that the auto-encoder here is irrelevant to the auto-encoder trained for the binary encoding vector $\vec{b}$ in Section 3.2. The AAE method outperforms the state-of-the-art embedding methods. For example, we apply exact K-NNS to retrieve products from an annotated private dataset of 100k products in 1300 categories. The performance is measured by precision, which calculates the proportion of retrieved products from the same product category as the query product. The precision of the exact 3-NNS using the AAE embedding is 0.744, while the corresponding precisions for vanilla fastText, universal sentence encoder, and Bert are 0.662, 0.651, and 0.615, respectively [7, 9, 16]. We omit more details due to the page limit. Instead, we choose to conduct experiments directly on the final search method in Section 4.

3.2 Binary Hashing via Denoising Autoencoder

In this subsection, we demonstrate how to obtain the binary encoding vector $\vec{b} = h(\vec{v})$. The basic idea is to use denoising auto-encoder to compress the embedding $\vec{v}$ into a lower dimension logit vector $\vec{z}$ and find an optimal thresholding vector $\vec{\theta}$ to convert $\vec{z}$ into binaries. The binary encoding schema originates from Hinton *et al.* [23]. Our work focuses on optimizing the threshold $\vec{\theta}$ for better retrieval quality and lower latency. To enhance the search method precision and latency, the following desired properties of the encoding are proposed:

(1) The encoding dimension should be low for efficiency.
(2) Tiny buckets should be avoided, otherwise NNS has to take place on too many buckets and latency issue will come up.
(3) Huge buckets should be avoided for ranking efficiency.
(4) The average bucket size should be controllable for different values of M.

Denoising auto-encoders are frequently used for extracting and composing robust features, where the neural network is fed with manually corrupted data to enhance feature quality and stability [31, 32]. The details of the network structure is well-known to the community, and hence we only demonstrate how to optimize the thresholding vector $\vec{\theta}$ in this paper. Our auto-encoder model consists of a two-layer encoder and a two-layer decoder, where the number of the encoded feature layer d' is set to 32 with sigmoid function as the activation. We use $\vec{z} = [z_1, z_2, \ldots, z_{32}]$ to denote the compressed feature from the encoding layer where $z_i \in (0, 1)$ $\forall 1 \le i \le 32$, and the final binary encoding is obtained via thresholding by $b_i = \mathbb{1}_{z_i \ge \theta_i}$.

Naively choosing $\theta_i = 0.5$ as in [23] will result in too many buckets with extreme sizes. Note that now the binary bucket $\mathcal{B} = \mathcal{B}(\vec{\theta})$ depends on the thresholding vector, and let $n_i(\vec{\theta}) = |\mathcal{B}_i(\vec{\theta})|$ be the bucket size. We define the following objective function:

$$L'(\vec{\theta}) := \sum_{i}^{|\mathcal{B}^*|} \left(\frac{\chi^4}{n_i^2(\vec{\theta})} + n_i^2(\vec{\theta}) \right), \tag{5}$$

where χ is a tunable parameter to balance the average bucket size and the number of buckets. For a single bucket, the summand in (5) is convex and has a unique minimum at $n_i = \chi$, and both huge and tiny n_i are naturally penalized. However, the objective (5) becomes non-convex and even discontinuous with respect to $\vec{\theta}$, making traditional optimization techniques unsuitable. To overcome the optimization difficulty, we adopt a continuous genetic algorithm [2], with population size 100, mutation rate 0.2, and in each generation 100 pairs are randomly selected for crossover. We iterate for 200 generations and choose the $\vec{\theta}$ with the smallest objective L' across all the generations. For our test product pool of 40 million products, the optimized $\vec{\theta}$ for $\chi = 100$ is able to increase the average bucket size from 1.78 via the vanilla semantic hashing to 57.32, and the largest bucket size decreases from 60k to 3.5k. Again, we omit the standalone evaluation of the proposed encoding method, and conduct experiments on the overall search method in Section 4.

4 EXPERIMENTS

In this section, we present a series of experiments to evaluate the proposed solution. We have built a Reference Product Service (RPS) using the AAE product embedding and semantic hashing approximate NNS. For a given query product, its embedding and binary encoding are computed in real-time, followed by the semantic hashing NNS, which is implemented using a CPU-GPU hybrid model. We first compare RPS with its peer services and present two quality metrics, namely the return rate and the precision at K, in Section

Table 1: Return rate at K, $\mathcal{R}_K(\mathcal{S}, Q_{\text{purchased}})$

K	RPS	Method-1	Method-2	Behavior-1	Behavior-2
1	0.922	0.293	0.865	0.643	0.742
3	0.902	0.293	0.748	0.528	0.688
5	0.893	0.149	0.637	0.475	0.657
10	0.888	0.0	0.408	0.405	0.606
50	0.775	0.0	0.0	0.234	0.415

Table 2: Return rate at K, $\mathcal{R}_K(\mathcal{S}, Q_{\text{general}})$

K	RPS	Method-1	Method-2	Behavior-1	Behavior-2
1	0.638	0.008	0.039	0.017	0.030
3	0.599	0.008	0.030	0.013	0.023
5	0.586	0.003	0.024	0.011	0.020
10	0.579	0.00	0.014	0.009	0.017
50	0.456	0.00	0.00	0.004	0.010

4.1 and 4.2, respectively. Then, as the semantic hashing NNS approximates the Exact NNS, we compare its performance with the Exact NNS in Section 4.3. Lastly, the complexity and computation efficiency are briefly discussed in Section 4.4.

4.1 Return Rate Test

For a given test pool of queries Q_{test}, let $\psi(\mathcal{S}, q)$ be the number of search results returned by $\mathcal{S}$ for the query product $q \in Q_{\text{test}}$. Then the return rate at K, denoted by $\mathcal{R}_K(\mathcal{S}, Q_{\text{test}})$ is defined as

$$\mathcal{R}_K(\mathcal{S}, Q_{\text{test}}) := \frac{1}{|Q_{\text{test}}|} \sum_{q \in Q_{\text{test}}} \mathbb{1}_{\psi(\mathcal{S}, q) \geq K}, \tag{6}$$

where $\mathbb{1}$ is the indicator function. The return rate estimates the ability to retrieve enough reference products for different query products. In reality, this ability is largely limited by product feature availability and the design of the search method. For example, a search method depending on customer browsing history will not work for a new product without views. However, recall that one of our goals of the reference products is to make it general so it is able to support different applications.

Two test sets are constructed, namely the purchased set and the general set. The purchased set consists of 10^6 randomly sampled products with purchase history. The general set consists of 10^6 products randomly sampled from a billion-level product pool. Products from the purchased set can be considered be higher quality. We compare the return rate of RPS with four existing product-to-product services. Two peer services highly depend on behavioral data (denoted by Behavior-1 and Behavior-2), and the other two depend on general product information (denoted by Method-1 and Method-2). RPS uses a product pool of 4×10^7 products with $d = 100$, $d' = 32$ and $M = 4000$. The related results are listed in Table 1 for the purchased query set and Table 2 for the general query set.

We acknowledge that we have no control on the quality or the size of the product pool used by each peer service, nor how the search method is implemented. We simply summarize the observed results as below. Our service returns enough search results for a larger proportion of query products. This advantage becomes more obvious for general products where all the peer services fail. Recall that the return rate estimates the applicability of a solution, better return rates indicate that our solution is able to support broader applications.

Table 3: Precision test in each product line

Product Line	Method-1	Method-2	RPS
Softline	0.582	0.572	0.865
Hardline	0.621	0.656	0.794
Consumable	0.673	0.705	0.827

4.2 Precision Test

The precision at K metric $\mathcal{P}_K$ is defined as the proportion of the top-K results that are indeed similar to the query based on human judgement. More specifically, let $\varphi(\mathcal{S}, p, K)$ be the number of positively annotated products from the top-K search results by method $\mathcal{S}$ for a query $q \in Q_{\text{test}}$, the precision at K metric $\mathcal{P}_K$ is defined as

$$\mathcal{P}_K(\mathcal{S}, Q_{\text{test}}) := \frac{1}{|Q_{\text{test}}|} \sum_{q \in Q_{\text{test}}} \varphi(\mathcal{F}, q, K). \tag{7}$$

We observe a sizable proportion of the search results from Behavior-1 and Behavior-2 is dominated by noise since both methods heavily depend on behavioral features, making them less competitive for the precision test. Thus, we only annotate the search results from Method-1, Method-2, and our own RPS. We set $K = 5$ and randomly sample 1500 products such that $\psi(\mathcal{S}, q) \geq 5$ for all the three methods. In order to better demonstrate the difference between the methods, the annotators are asked to use strict criterion in terms of similarity and to label at least 3 products as negative out of the 15 products pooled from the top-5 results from each of the three method. The precision test results are shown in Table 3 where RPS outperforms the other two baseline peers. Note that the results should be viewed for comparison purpose only and the actual precision is higher for all the methods because of the strict criterion.

4.3 Approximate KNN Recall Test

In addition to the above search quality test, we also conduct the recall test for completeness. The recall rate measures the proportion of actual positives that are included in the search results, which is generally not practical since the ground truth is unknown. Instead, we follow the convention to compute the recall against the exact NNS method and compare it with four other open-source approximate NNS methods. Non-Metric Space Library (NMSLIB) [22], Approximate Nearest Neighbors Oh Yeah (ANNOY) [10], and Fast Library for Approximate Nearest Neighbors (FLANN) are tree-based approaches, while Iterative Expanding Hashing is a hashing type approach [6, 11, 20, 21].

Since we require the exact KNN results to be the ground truth, we use a smaller dataset for the recall test: a set of 2 million products generated randomly. For each method, we calculate the proportion of the top-K search results that actually hit the top-K products from exact KNN search. Recall values are presented in Table 4 for K

Table 4: Recall-at-K test

K	RPS	NMSLIB	FLANN	ANNOY	IEH
1	97.24	90.13	90.59	91.35	95.14
5	95.82	93.42	85.57	85.94	90.88
10	94.85	92.67	79.10	82.62	87.65
100	90.72	61.90	58.64	66.90	74.21

varying from 1 to 100. Note that the query itself is always excluded from the search results.

Table 4 shows that our approximate KNN based on binary encoding achieves satisfactory recall rates. In addition, we would like to highlight its flexibility. The efficiency-quality parameter M can be adjusted online without preprocessing to balance efficiency and quality, while most of other approximate KNN algorithms can not. It also allows easy modification of product pool in realtime. For example, adding new embedding and encoding vectors are more easily handled than training a new tree structure.

4.4 Latency

In this subsection, we briefly discuss the efficiency and the latency of our search method. In general, the combined embedding and encoding latency for a single query is under 1ms. With a single AWS machine of type p2.xlarge (equipped with an NVIDIA K80 GPU), for a product pool of 40 million products and the subset size $M = 4000$, the average search latency is under 6ms, which is in the same scale as the general network latency. Thus, the proposed method is able to efficiently support general business applications.

5 DISCUSSION AND CONCLUSION

In this paper, we propose a search method to surface a set of reference products. The product catalog information is transformed into a high-quality embedding of low dimension via a novel attention auto-encoder neural network, and the embedding is further coupled with a binary encoding vector for high quality vector search at scale. We conduct extensive experiments to evaluation the return rate, the precision, and the recall rate of the proposed method, and compare them with peer methods. Since our method is able to yield a satisfactory number of high-quality results for most of the query products, the reference products are readily consumable for various business ranking models to support applications like pricing, substitution, and recommendation. We believe such an acceleration to support multiple applications will bring fundamental values to the e-commerce business, and invite more future works on the algorithms and applications of the reference product set.

REFERENCES

[1] David W Aha, Dennis Kibler, and Marc K Albert. 1991. Instance-based learning algorithms. *Machine learning* 6, 1 (1991), 37–66.

[2] Christine M Anderson-Cook. 2005. Practical genetic algorithms.

[3] Alex Auvolat, Sarath Chandar, Pascal Vincent, Hugo Larochelle, and Yoshua Bengio. 2015. Clustering is efficient for approximate maximum inner product search. *arXiv:1507.05910* (2015).

[4] Yoram Bachrach, Yehuda Finkelstein, Ran Gilad-Bachrach, Liran Katzir, Noam Koenigstein, Nir Nice, and Ulrich Paquet. 2014. Speeding up the xbox recommender system using a euclidean transformation for inner-product spaces. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 257–264.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[6] Leonid Boytsov and Bilegsaikhan Naidan. 2013. Engineering efficient and effective non-metric space library. In *International Conference on Similarity Search and Applications*. Springer, 280–293.

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv:1803.11175* (2018).

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018).

[10] Bernhardsson Erik. 2016. Approximate Nearest Neigbors OnYeah (Annoy). https://github.com/spotify/annoy.

[11] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. 2013. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2916–2929.

[12] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine.* " O'Reilly Media, Inc.".

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Gisli R Hjaltason and Hanan Samet. 2003. Index-driven similarity search in metric spaces (survey article). *ACM Transactions on Database Systems (TODS)* 28, 4 (2003), 517–580.

[15] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2011. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2011), 117–128.

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv:1607.01759* (2016).

[17] Noam Koenigstein, Parikshit Ram, and Yuval Shavitt. 2012. Efficient retrieval of recommendations in a matrix factorization framework. In *International conference on Information and knowledge management*. ACM, 535–544.

[18] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.

[19] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 1 (2003), 76–80.

[20] Yury A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[21] Marius Muja and David G Lowe. 2014. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 11 (2014), 2227–2240.

[22] Bilegsaikhan Naidan, Leonid Boytsov, Malkov Yury, Novak David, and Frederickson Ben. 2016. Non-Metric Space Library (NMSLIB). https://github.com/nmslib/nmslib.

[23] Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning* 50, 7 (2009), 969–978.

[24] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press.

[25] Thomas Seidl and Hans-Peter Kriegel. 1998. Optimal multi-step k-nearest neighbor search. In *ACM Sigmod Record*, Vol. 27. ACM, 154–165.

[26] Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems*. 2321–2329.

[27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014).

[28] David Smiley, Eric Pugh, Kranti Parisa, and Matt Mitchell. 2015. *Apache Solr enterprise search server*. Packt Publishing Ltd.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[30] Sudheendra Vijayanarasimhan, Jonathon Shlens, Rajat Monga, and Jay Yagnik. 2014. Deep networks with large output spaces. *arXiv:1412.7479* (2014).

[31] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. ACM, 1096–1103.

[32] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, Dec (2010), 3371–3408.

[33] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. 2014. Hashing for similarity search: A survey. *arXiv:1408.2927* (2014).
[34] Jason Weston, Samy Bengio, and Nicolas Usunier. [n. d.]. Wsabie: Scaling up to large vocabulary image annotation.
[35] D Randall Wilson and Tony R Martinez. 2000. Reduction techniques for instance-based learning algorithms. *Machine learning* 38, 3 (2000), 257–286.
[36] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. 2014. Supervised hashing for image retrieval via image representation learning.

ProductNet: a Collection of High-Quality Datasets for Product Representation Learning

Chu Wang
Amazon.com
chuwang@amazon.com

Lei Tang
Amazon.com
leitang@amazon.com

Yang Lu
Amazon.com
ylumzn@amazon.com

Shujun Bian
Amazon.com
sjbian@amazon.com

Hirohisa Fujita
Amazon.com
hirohisf@amazon.com

Da Zhang
Amazon.com
dazh@amazon.com

Zuohua Zhang
Amazon.com
zhzhang@amazon.com

Yongning Wu
Amazon.com
yongning@amazon.com

ABSTRACT

ProductNet is a collection of high-quality product datasets for better product understanding. Motivated by ImageNet, ProductNet aims at supporting product representation learning by curating product datasets of high quality with properly chosen taxonomy. In this paper, the two goals of building high-quality product datasets and learning product representation support each other in an iterative fashion: the product embedding is obtained via a multi-modal deep neural network (master model) designed to leverage product image and catalog information; and in return, the embedding is utilized via active learning (local model) to vastly accelerate the annotation process. For the labeled data, the proposed master model yields high categorization accuracy (94.7% top-1 accuracy for 1240 classes), which can be used as search indices, partition keys, and input features for machine learning models. The product embedding, as well as the fined-tuned master model for a specific business task, can also be used for various transfer learning tasks.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Learning latent representations**; **Active learning settings**.

KEYWORDS

Representation Learning, Multi-Modal Learning, Deep Learning, Dataset Construction, Active Learning

ACM Reference Format:
Chu Wang, Lei Tang, Yang Lu, Shujun Bian, Hirohisa Fujita, Da Zhang, Zuohua Zhang, and Yongning Wu. 2019. ProductNet: a Collection of High-Quality Datasets for Product Representation Learning. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316607

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316607

1 INTRODUCTION

E-Commerce retail is all about products, be it physical goods or digital content. Product representation learning is the key to enhancing search and discovery experiences for customers, and product management for the backend systems. Compared to raw attributes like image and catalog information, the product representation has advantages of maintaining product semantic information and being in a compact form to facilitate modeling and algorithm designs. Therefore, the major problem becomes how to build high-quality product representation with satisfactory transferability. In order to learn how customers understand products, motivated by the success of ImageNet, we took a data first and quality first approach to facilitate product representation learning, by developing ProductNet, a collection of high-quality labeled product datasets.

Building high-quality datasets and categorizing products into thousands of classes are difficult; developing product embedding with satisfactory transferability to incorporate product image and text information is even more challenging. To achieve these two goals, instead of utilizing large-volume noisy data, we choose to use small-volume, high-quality *gold datasets*. That being said, we explore the quality dimension of the data before the volume dimension by working on a classification task via human annotation. The two goals of product categorization and representation learning support each other in an iterative fashion within our work: the product embedding is built based on deep classifiers for the labeled dataset; the embedding itself, in return, is utilized in an active learning framework to enhance the labeling speed and quality.

Note that the goal of annotation is not to cover billions of products. Instead, we focus on a subset of high-quality products for the representation learning purpose. In particular, we aim at the diversity and representativeness of the products. Being representative, the labeled data can be used as reference products to power product search, pricing, and other business applications. Being diverse, the models are able to achieve strong generalization ability for unlabeled data, and the product embedding is also able to represent richer information. A carefully chosen taxonomy is important to reduce annotation ambiguities and mistakes. We adopt a function-based product taxonomy for the ProductNet construction. As for now, we have populated 3900 categories of non-media products,

with roughly 40-60 products for each category. For either categorization or representation learning, we need to handle products with noisy attributes, missing fields, or even manually altered product information. Combining different attributes of products is one way to alleviate these potential problems and to achieve better robustness. We utilize multi-modal learning to incorporate information from different fields into the feature embeddings and to deal with the issue of noisy or missing product attributes.

We are able to achieve 94.7% top-1 accuracy on our product categorization task (1240 classes). The high-accuracy comes from both the multi-modal model and the curated dataset with carefully chosen taxonomy. Such high accuracy opens the door for the predicted categorization to be used as search indices and partition keys. We hope carefully curated product datasets, like ProductNet, will lead a better way for product categorization and product representation learning. Before going into details, we first discuss the challenges of building high-quality datasets and related works.

Challenges. Building a high quality dataset is never a trivial task and it faces several challenges: i) the annotation quality to provide correct labels; ii) the annotation quantity to support a large scale dataset; iii) noisy or missing fields (e.g. misplaced product description or empty product image); iv) distinct data format from different fields like categorical attribute, free-form text, and images.

Related work. ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which collects more than 1.2 million varied images, has inspired the development of a series of deep learning models [3, 7, 13, 14].The last hidden layer signals of deep models are widely used as image feature embeddings for transfer learning [2]; the upper-layers of deep models can also be fine-tuned for different tasks. Natural language processing benefits largely from the introduction of word2vec [11] and its extension to sentences, documents, and sub-word information [1, 8]. In addition to ImageNet, there are various datasets which contribute significantly to model development and real-world applications [9, 10].

2 DATASET CONSTRUCTION

In this section, we demonstrate and discuss how we construct ProductNet datasets. A naive way for obtaining product labels is directly via human annotation: given a *candidate pool*, each product is tagged with a category name by human. This way of labeling is unbearably expensive as long as the number of categories is moderately large. Furthermore, since the annotation involves choosing the correct category out of many, background knowledge is required for the labeler and mistakes can not be avoided.

Even though human annotation itself is irreplaceable, we proceed in an intelligent way in order for higher efficiency and data quality. The entire workflow of ProductNet dataset construction is based on an iterative loop of human annotation and representation learning. Before going into a detailed demonstration, we would like to highlight the importance of taxonomy in the construction of labeled datasets. Product understanding and representation learning can largely benefit from a taxonomy that approximates the intrinsic data distribution well. There are different types of taxonomies, including function-based, subject-based, and organization-based ones. To avoid ambiguity and improve product embedding, we prefer the function-based taxonomy for non-media products.

2.1 Human annotation and local model

We annotate products category by category. In this way, an annotator only needs to make a binary decision on whether a product belongs to the current category (positive) or not (negative). In the meantime, only the background knowledge for the current category is required. We maintain a set of *local models* to provide suggestions of products to be annotated based on available labels. The local model can be a KNN search engine or a keyword-based search engine. By searching similar products from the pool, the engine is able to provide relevant candidate products.

Binary classifiers based on active learning mechanism are more helpful as local models. The binary classifiers are built on and refined by both the positive and negative samples. With at least one positive product and one negative product, a binary classifier is trained to initialize the active learning process, which is then able to provide positive and negative suggestions for labeling. More importantly, the algorithm asks for labels of ambiguous candidates in order to improve the local model itself, thus the quality of positive and negative suggestions becomes better. To promote better variety and retrievability of the products, we use KNN search, keyword search, and active learning methods on generalized linear models, in a mixed way. A screenshot of ProductNet annotation portal is demonstrated in Figure 1, and more discussion of the active learning sampling is in Section 3.1

2.2 Representation learning and master model

When we have labeled positive data from multiple categories, a multi-modal classification neural network is trained for the goal of representation learning. We name this deep neural network the *master model*. Based on the gold dataset, we train the master model and extract the last hidden layer signals as product feature embeddings. The embeddings are then fed back to the active learning module for another round of annotation. In addition to the feature embeddings produced by the master model, we also get the product categorization. As long as the master model yields high accuracy, its prediction also provides accurate product candidate for annotation.

Formally, let $c = (1, \cdots, C)$ be the class label of a product, x be the input data representing both text and image data sources. We learn a conditional distribution $p(c|x)$ defined as

$$p(c|x) = \frac{\exp\{f_c(x; w) + b_c\}}{\sum_{c=1}^{C} \exp\{f_c(x; w) + b_c\}} \tag{1}$$

where $f_c(x; w)$ is the scoring function from class c, b_c is the bias term, and w collects all trainable weights. The scoring function $f_c(x; w)$ fuses both text and image information computed by neural networks. The complete structure of our master model is shown in Figure 2. We discuss our design of the master model from different aspects in the following several sessions.

We adopt Inception-v4 as our image model [14]. When the image is missing for a product, a zero image is used instead. The output of image model is a 1536-dimensional vector from mixed_7d layer from Inception-v4 for model fusion. There are seven fields of catalog data that we have considered: product title, product description, bullet points, brand, and three types of keywords. In spite of missing fields and varying qualities, the multi-modal deep neural network is able to adjust the relative weights for better results. We choose deep

Figure 1: Screenshot of ProductNet annotation portal.

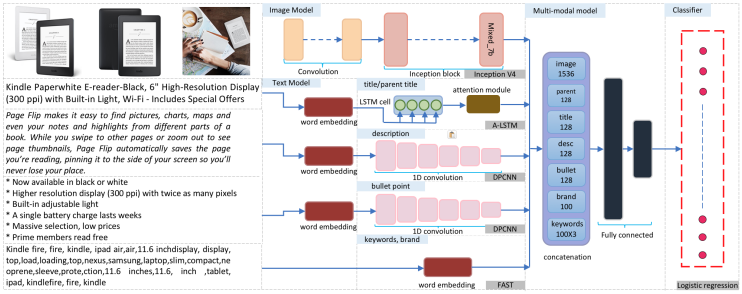


Figure 2: Master model structure: multi-modal deep neural network.

pyramid convolutional neural network (DPCNN) for description and bullet points, attention-LSTM (A-LSTM) for product title, and vanilla fasText (FAST) for the brand and keywords [5, 6, 12, 15]. The models for each field are selected by running experiments on each field alone and adopting the one with the best performance. A multi-modal model builds a joint representation of different data sources and they could reinforce each other. We learn a fused model integrating both text signals and image signals. The final feature vector is composed by concatenating the outputs from image and text models with two fully-connected layers. Note that the dimension of the feature embeddings is adjustable, which is helpful when low-dimension embedding is needed for higher efficiency.

2.3 Iterative construction of datasets

The representation learning module and the human annotation module operate iteratively. For any given candidate pool and taxonomy, initial product embeddings are required to start the human annotation process. The initial embeddings can be obtained via pre-trained models. For example, we choose Inception-v4 trained on ImageNet for image data, and fastText model for text data.

Note that it is not necessary to use high quality embeddings initially. The initial feature embeddings will support retrieval methods like active learning for the first round of annotation, and then the labels are used for training the master model and producing a new version of product embeddings. With the new embeddings, we conduct another human annotation stage. But this time, the speed and quality are further boosted by the embeddings.

One of the advantages of our iterative construction is the ability of local adjustment. Note that if we want to apply active learning directly to the master model, the newly labeled data will have very small impact on the model because the number of classes is large. With the local models, on the other hand, we are able to conduct local adjustment precisely and only focus on the categories of interest. Furthermore, we frequently mark the wrongly labeled data by the master model as local negatives, so that further annotated data will direct the master model to the more accurate direction.

3 EXPERIMENTS

In this section, we demonstrate our experiment results. As a proof of concept, we adopt the Google Taxonomy [4], which partitions products based on their functionalities in a tree structure. As the paper submitted, our labeled dataset covers 3900 leaf nodes with 178k products. Most of products have both text and image information. Each leaf node is labeled by one labeler and verified by one auditor independently to reduce labeling errors. At each leaf node, we also maintain a similar number of negative samples which are most likely to be confused with positive ones for labeling reference and training the local model. To evaluate our dataset, we first exhibit some qualitative results regarding the acceleration of labeling by active learning, then evaluate the accuracy of mater model prediction, and discuss how combining the two accelerates the dataset construction in the end.

3.1 Active learning recommendation

We demonstrate how our local model and active learning technique accelerate the labeling process. The product retrieval module via active learning and other methods is located in the right bottom of Figure 1, Currently, we support random sampling, keyword search,

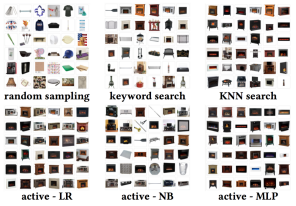


Figure 3: Comparison of different sampling methods.

KNN search, ad-hoc input, and active learning based on logistic regression (LR), naive Bayes (NB), or multi-layer perceptron (MLP). As an example, we demonstrate the sampled products for the leaf node "Fireplaces" in Figure 3. It can be observed that random sampling can hardly yield any fireplace from the candidate pool; the keyword search is able to get several fireplaces while most of the search results are irrelevant; active learning and KNN search are able to provide meaningful recommendations to enhance the labeling speed by a significant amount. Furthermore, the active learning methods and KNN search are able to provide different types of fireplaces to further enhance product variety.

3.2 Master model classification

We test the master model on a subset of 1240 leaf nodes, each having 40 products. For each category, 32 products (80%) are randomly chosen as the training set, and the rest as test set. The classification accuracy is high, indicating the powerful generalization ability of our multi-modal model. The quality of our dataset also contributes to the high accuracy, as any defects within categories will lower the prediction accuracy. The training process takes 10 hours on a single AWS p3.2xlarge GPU instance for 50 epochs to achieve 94.7% accuracy and become stabilized. We demonstrate the performance of the master model with image and text sources (master-IT) and show its improvement with non-multimodal methods like Inception-v4, bag-of-words (BOW), and text-only master model (master-T) in Table 1. Though the high classification accuracy demonstrates the advantage of the master model structure, we would like to highlight the contribution of the dataset being of high quality. With a properly designed taxonomy and accurately annotated products, the demand for sophisticated models decreases, while the model transferability increases.

Table 1: Master model classification accuracy (percentile).

	Inception-v4	BOW	master-T	master-IT
top-1 acc	69.4	83.1	92.6	94.7
top-3 acc	85.5	90.7	97.8	98.2
top-5 acc	90.3	93.1	98.5	99.1

3.3 Annotation Acceleration

We test the acceleration of the active learning retrieval for the annotation, and find that a normal annotator with general background knowledge is able to label 100 positive data points for each leaf node within 30 minutes. Compared to vanilla labeling which takes 30 minutes to find 5 positive data points for a leaf node, the acceleration factor is roughly 20. Furthermore, after the master model is trained, the products from the annotation pool will be given a recommended category by the master model so that the annotation process can be further accelerated by verifying that category label, and we estimate the annotation acceleration factor via master model is 80 compared to vanilla labeling.

4 CONCLUSION

We have introduced ProductNet, a collection of high-quality product datasets for better product understanding. Our framework is a fast and reliable way of constructing product labels and building high-quality datasets. The master model is able to provide business acceptable labels for product listings, product indexing, and partition keys; and the product embedding obtained can support various product modeling tasks and business applications. The experiments verify our initiative that a dataset of high-quality is able to foster high-quality product embeddings.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018).
[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE conference on computer vision and pattern recognition*. 580–587.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[4] Google Inc. 2015. Google Product Taxonomy. https://www.google.com/basepages/producttype/taxonomy.en-US.txt.
[5] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 562–570.
[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv:1607.01759* (2016).
[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*. 1097–1105.
[8] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.
[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[10] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 165–172.
[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[12] Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Conference of the International Speech Communication Association*.
[13] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014).
[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In *AAAI*, Vol. 4. 12.
[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762* (2017).

Emoji2019 Workshop Chairs' Welcome

It is our great pleasure to welcome you to *Emoji2019: The 2nd International Workshop on Emoji Understanding and Applications in Social Media at WWW 2019*. With the rise of social media, emoji has become a popular and an indispensable form of online communication. Emojis have been widely used to enhance the sentiment, emotion, and sarcasm expressed online. As emojis offer a primary method for more expressive online communication, it is essential to investigate how they can be used to improve computer-mediated communication. However, the ability to automatically process, derive meaning, and interpret text fused with emoji is challenged by the pictorial nature of emoji, the fact that the same emoji may be used in different contexts to express different meanings, and that in different cultures and communities over the world the same emoji representation may be interpreted differently. This makes it especially difficult to apply traditional Natural Language Processing techniques to analyze them. The Emoji2019 workshop seeks to provide a forum to bring together researchers and practitioners from both academia and industry in the areas of computer science, social network analysis and mining, natural language processing, computational linguistics, human-computer interaction, and computational social sciences to discuss high-quality research and emerging applications, to exchange ideas and experience, and to identify new opportunities for collaboration towards solving the above challenges.

We received a total of 10 submissions out of which, 6 were accepted to present at the workshop. The accepted papers cover a broad range of topics including emotional valance of emoji and its variation across genders, emoji prediction, emoji sequence mining, emoji as location markers, and emoji-first communication among others. The technical program is complemented and enhanced with a keynote speech and three invited talks. Professor Qiaozhu Mei of the University of Michigan, USA, will give a keynote speech on "Decoding the New World Language: Analyzing the Popularity, Roles, and Utility of Emojis", where he would emphasize how emojis have surpassed its original intentions to become a new form of communication and the unique opportunities this affords for multidisciplinary research. Jennifer 8. Lee, Founder of Emojination, will be talking about her findings on the resemblance of Chinese characters to emoji. Jennifer Daniel, Creative Director (emoji) at Google Inc., will discuss the importance of having gender inclusive emoji designs and the challenges in designing them. Benjamin Weissman of the University of Illinois at Urbana-Champaign, USA, will share his findings on the similarities between the way the brain reacts to emoji and to irony.

We believe that the wide variety of research topics covered in Emoji2019 workshop program provides a forum that could lead to thought-provoking discussions among the researchers from multiple disciplines who will be attending the workshop. We hope this year's workshop would lead to new collaborations that would help further emoji research and applications.

Amit Sheth
Kno.e.sis Center, Wright State University, USA

Horacio Saggion
Universitat Pompeu Fabra, Spain

Sanjaya Wijeratne
Kno.e.sis Center, Wright State University, USA

THE WEB CONFERENCE

Emoji2019 Workshop Organization

Organization Committee:
Amit Sheth *(Kno.e.sis Center, Wright State University, USA)*
Horacio Saggion *(Universitat Pompeu Fabra, Spain)*
Sanjaya Wijeratne *(Kno.e.sis Center, Wright State University, USA)*

Program Committee:
Alexander Robertson *(University of Edinburgh, United Kingdom)*
Anurag Illendula *(Indian Institute of Technology – Kharagpur, India)*
Derek Doran *(Kno.e.sis Center, Wright State University, USA)*
Erik Cambria *(Nanyang Technological University, Singapore)*
Fan Yang *(Wright State University, USA)*
Francesco Barbieri *(Universitat Pompeu Fabra, Spain)*
Francesco Ronzano *(Universitat Pompeu Fabra, Spain)*
Hannah Miller Hillberg *(University of Wisconsin – Oshkosh, USA)*
Jacob Thebault-Spieker *(Virginia Polytechnic Institute and State University, USA)*
Jing Ge *(University of California, Berkeley, USA)*
Lakshika Balasuriya *(Gracenote Inc., USA)*
Ning Xie *(Wright State University, USA)*
Petra Kralj Novak *(Jožef Stefan Institute, Slovenia)*
Sarasi Lalithsena *(IBM Almaden Research Center, USA)*
Sashank Santhanam *(University of North Carolina at Charlotte, USA)*
Sujan Perera *(IBM Almaden Research Center, USA)*
Tianran Hu *(University of Rochester, USA)*
Tyler Schnoebelen *(Integrate.ai, Canada)*
Umashanthi Pavalanathan *(Georgia Institute of Technology, USA)*
Wei Ai *(University of Michigan, USA)*

Decoding the New World Language: Analyzing the Popularity, Roles, and Utility of Emojis

Qiaozhu Mei
University of Michigan
Ann Arbor, Michigan
qmei@umich.edu

ABSTRACT

Emojis have quickly become a universal language that is used by worldwide users, for everyday tasks, across language barriers, and in different apps and platforms. The prevalence of emojis has quickly attracted great attentions from various research communities such as natural language processing, Web mining, ubiquitous computing, and human-computer interaction, as well as other disciplines including social science, arts, psychology, and linguistics.

This talk summarizes the recent efforts made by my research group and our collaborators on analyzing large-scale emoji data. The usage of emojis by worldwide users presents interesting commonality as well as divergence. In our analysis of emoji usage by millions of smartphone users in 212 countries, we show that the different preferences and usage of emojis provide rich signals for understanding the cultural differences of Internet users, which correlate with the Hofstede's cultural dimensions [4].

Emojis play different roles when used alongside text. Through jointly learning the embeddings and topological structures of words and emojis, we reveal that emojis present both complementary and supplementary relations to words. Based on the structural properties of emojis in the semantic spaces, we are able to untangle several factors behind the popularity of emojis [1].

This talk also highlights the utility of emojis. In general, emojis have been used by Internet users as text supplements to describe objects and situations, express sentiments, or express humor and sarcasm; they are also used as communication tools to attract attention, adjust tones, or establish personal relationships. The benefit of using emojis goes beyond these intentions. In particular, we show that including emojis in the description of an issue report on GitHub results in the issue being responded to by more users and resolved sooner.

Large-scale emoji data can also be utilized by AI systems to improve the quality of Web mining services. In particular, a smart machine learning system can infer the latent topics, sentiments, and even demographic information of users based on how they use emojis online. Our analysis reveals a considerable difference between female and male users of emojis, which is big enough for a machine learning algorithm to accurately predict the gender of a user. In Web services that are customized for gender groups, gender inference models built upon emojis can complement those based on text or behavioral traces with fewer privacy concerns [2].

Emojis can be also used as an instrument to bridge Web mining tasks across language barriers, especially to transfer sentiment knowledge from a language with rich training labels (e.g., English) to languages that have been difficult for advanced natural language processing tasks [3]. Through this bridge, developers of AI systems and Web services are able to reduce the inequality in the quality of services received by the international users that has been caused by the imbalance of available human annotations in different languages.

In general, emojis have evolved from visual ideograms to a brand-new world language in the era of AI and a new Web. The popularity, roles, and utility of emojis have all gone beyond people's original intentions, which have created a huge opportunity for future research that calls for joint efforts from multiple disciplines.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Natural language processing**.

KEYWORDS

emojis; text mining; natural language processing

ACM Reference Format:
Qiaozhu Mei. 2019. Decoding the New World Language: Analyzing the Popularity, Roles, and Utility of Emojis. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316541

BIOGRAPHY

Qiaozhu Mei is an associate professor in the School of Information and the Department of Electrical Engineering and Computer Science at the University of Michigan. His research focuses on large-scale data mining, machine learning, information retrieval, and natural language processing, with broad applications to social media, Web search, and health informatics. Qiaozhu is an ACM distinguished member (2017) and a recipient of the NSF Career Award (2011). His work has received multiple best paper awards and nominations at ICML, KDD, WWW, WSDM, and other major conferences. He has served as the General Co-Chair of SIGIR 2018 and on the editorial boards of multiple journals such as JMLR, TOIS, and TWEB. He is the founding director of the new master degree of applied data science at the University of Michigan.

Qiaozhu received his PhD degree in computer science from the University of Illinois at Urbana-Champaign and his Bachelor of science degree from Peking University.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316541

REFERENCES

[1] Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *Eleventh International AAAI Conference on Web and Social Media*.

[2] Zhenpeng Chen, Xuan Lu, Wei Ai, Huoran Li, Qiaozhu Mei, and Xuanzhe Liu. 2018. Through a gender lens: learning usage patterns of emojis from large-scale android users. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 763–772.

[3] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-Lingual sentiment classification. In *Proceedings of the 2019 World Wide Web Conference*.

[4] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 770–780.

Assessing the Meaning of Emojis for Emotional Awareness - A Pilot Study

Wesley Brants
Eindhoven University of Technology
Eindhoven, The Netherlands
w.m.g.brants@student.tue.nl

Bonita Sharif
University of Nebraska - Lincoln
Lincoln, Nebraska, USA
bsharif@unl.edu

Alexander Serebrenik
Eindhoven University of Technology
Eindhoven, The Netherlands
a.serebrenik@tue.nl

ABSTRACT

Emojis are increasingly being used in today's social communication - both formally in team messaging systems as well as informally via text messages on phones. Besides being used in social communication, emojis might also be a suitable mechanism for emotion (self-)assessment. Indeed, emojis can be expected to be familiar to people of different social groups and do not depend on the mastery of a specific language. However, emojis could be interpreted very differently from their actual intent. In order to determine whether people interpret emojis (specific to emotional states) in a consistent manner, we conducted an online survey on nine emojis with 386 people. The results show that the emojis representing anger, sadness, joy, surprise, and neutral state are interpreted as they were intended, independent of age and gender. Interpretations of other emojis such as *Unamused Face* and *Face Screaming in Fear* depend on age, and thus are not as useful for probing for emotion in a study setting unless all participants belong to the same age category. The *Face with Rolling Eyes* emoji is interpreted differently by gender and finally the *Nauseated Face* emoji resulted in no conclusive interpretation.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → **Gender**; **Age**.

KEYWORDS

emoji interpretation; pilot study; emotional awareness, gender, age

ACM Reference Format:
Wesley Brants, Bonita Sharif, and Alexander Serebrenik. 2019. Assessing the Meaning of Emojis for Emotional Awareness - A Pilot Study. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316550

1 INTRODUCTION

Emotions and their awareness can have a big impact on the success or failure of a project, both from a manager's [3] and employee's [7] perspective. This holds for collaborative projects in general, including software development projects [28]. The overall goal of our research is to provide members of collaborative teams with ways of indicating their emotion to other team members; however, to reduce the negative impact of interruption, soliciting emotion reporting should be accurate and quick. Both emojis and words can be used to convey the current mood of a subject, however emojis (and images in general) are more easy to interpret and recall than written words. This effect is known as the "picture superiority effect" and is due to dual-encoding of the pictures in the brain as an image and as a verbal trace [5, 22]. Moreover, use of graphic representations also works better for people who are dyslexic or otherwise aphasic. Lastly, there is no language barrier to consider.

Emojis are popular and used frequently used in many different settings [17]. However, due to the different representation of emojis on each platform and device [18], there could be a difference in perception of sentiment conveyed by an emoji. Moreover, sentiment might be misconstrued due to the same visual representation being interpreted differently [18]. Thus using emojis as indicator of one's emotional state might result in false conclusions if a person interprets an emoji differently as was intended by the researcher.

In this work, we analyze differences in interpretation of emotion-related emojis. Specifically, we consider *the impact of gender and age on the interpretation of emotion-related emojis* as gender and age have been shown to affect the use of emojis [21, 27]. We do this as part of assessing the suitability of emojis in probing for emotional awareness state.

Our results summarized in Table 1 show that interpretation of two of the nine emotion-related emojis are age-dependent, and one is gender-dependent. This means that using these three emojis to probe the emotional state of subjects should only be done in presence of gender- or age-homogeneous groups.

2 RELATED WORK

Measuring the emotional state of subjects can be done via several techniques. One category of techniques observes the subject, an activity or analyzes an artifact to infer the emotion of the participant. An example of this category of techniques is a lexicon approach, in which the emotional state is deducted from the analysis of words and sentences in which emotional words (i.e. love, shame, etc.) are given an emotional weight. There are multiple tools [13] that can be used for this such as SentiStrength [25], Stanford NLP sentiment analyser [24] and NLTK [16]. Moreover, importance of domain-specific aspects has been recognized, e.g., for software engineering [14] triggering development of a series of sentiment analysis tools targeting this particular domain. Another method of inferring the emotional state of a subject is the use biometric data such as pupil size, heart rate [8, 20] and galvanic skin response.

A second category of techniques to capture the emotional state is to explicitly ask the subject to indicate their his or her emotional

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316550

Table 1: Is interpretation of emojis independent from gender and age?

Description	Emoji	Gender	Age
Pouting Face	😡	✓	✓
Unamused Face	😒	✓	✗
Nauseated Face	🤢	✓	✓
Face Screaming in Fear	😱	✓	✗
Face with Rolling Eyes	🙄	✗	✓
Crying Face	😢	✓	✓
Face with Tears of Joy	😂	✓	✓
Face with Open Mouth	😮	✓	✓
Neutral Face	😐	✓	✓

state. This can for example be done prior to the task at hand [12]. There are several ways to do so; one can use a scale to indicate to what extent the participant is feeling an emotion, or one can do this via a standardized questionnaire such as Profile of Mood States (POMS) [15] or Positive and Negative Affect Schedule (PANAS) [26]. Other techniques besides these interviews and questionnaires are the use of film and pictures (e.g., the International Affective Picture System (IAPS)) [4]. Due to the picture superiority effect, using visual cues is not a strange choice. Combined with the familiarity of emojis and all benefits over written words it gives as explained before, emojis seem a logical choice for probing emotional state of the participant.

We report on studies done on the use of emoticons and emojis. Previous research shows that females are more accurate than males in judging emotional meaning from nonverbal cues [9]. Hudson et al. report on a study of emoticons showing that females are more jealous on Facebook, but when a winking emoticon ;) was present, males were more jealous [10]. Also, studies have shown that females have given more emotional dimensions to emoticons than males, meaning they interpret emoticons in more ways than do males [27]. A study on the use of emoticons in Japan hypothesized that females use emoticons more frequently than males. Also, younger people would use emoticons more frequently as compared to older people [21].

Emojis are quite different from emoticons. Emoticons are strings of characters representing some meaning such as sad or happy; emojis are images with little details conveying a more subtle meaning [18]. In the current literature, emojis have not yet been researched to the same extent as emoticons.

It has been shown that males and females use emojis differently and give them different sentiments. Also, there is a difference in frequency of emoji use depending on the time of the day, week and year [2]. For example, the emojis representing Christmas trees are used more often in December. The use of emojis has been shown to be different between males and females where different emojis are used at different times of the day, week and year as well as using

Table 2: Age Distribution of 386 respondents

Age	Participants	Age	Participants
18-23	67	36-41	55
24-29	110	≥ 42	79
30-35	72	did not answer	3

different sentiments for emojis. Age might also be a factor that influences usage as different generations communicate in different ways. Nishimura et al. hypothesized that since younger generations have more experience with emojis, they would use them more frequently [21]. This hypothesis was not supported by the results.

Research on the usage of emojis in food quality surveys has shown that emoji interpretation and frequency of use was not influenced by gender. However, when emojis are used to measure product-related emotional associations, age-related differences exist [11]. Our research has a different goal of determining whether emojis are suitable to probe emotional awareness, rather than rating a product.

3 STUDY OVERVIEW

The goal of this study is to infer whether emotion-related emojis are indeed a suitable way of determining the emotional state of a participant in a research study. This can only be done if the interpretations of emojis do not vary too much. To generalize the application of the results of this work, we survey a diverse group of participants. Since the usage of emojis is influenced by gender and age we ask the following research question *Does gender or age influence the interpretation of an emoji?* Based on the answer obtained we can also determine what emojis are suited for determining emotional awareness in gender and age diverse target groups.

We use nine emotion-related emojis for this study as shown in Table 1. Emojis are presented according to the Unicode style used by Apple Inc. These emojis were chosen because they are most commonly associated with emotions. Seeing whether interpretations are different in different age and gender categories is not entirely new. For some emojis the interpretation and usage of these emojis in questionnaires have been researched [11]. However, we perform this research with a different goal in mind; namely if they are usable in research in which emotional state is important. Moreover, we partly use other emojis which seem fit for this kind of goal. In order to answer the research question, we created an online survey and advertised it on Facebook and user groups. The participants were first asked to indicate their gender and age. Then, they were presented with 9 emojis in Table 1 and were asked to answer the question "What emotion is this emoji displaying?". The participants were allowed to provide a free-form answer in any language, and were not restricted to any length in response. We did this to obtain information that is as rich as possible [6]. A total of 389 people participated of which 3 responses were discarded.

4 DATA LABELING

The survey resulted in many different kinds of responses, varying from single words ("Bored", "Anger") to descriptive sentences ("it's like well I prefer not to comment about it"). We did not remove

Table 3: Clustered contingency table for gender and the Face With Rolling Eyes emoji. The Other category stands for Fear, Trust, Joy and Anticipation

	Anger	Disgust	Sadness	Surprise	Lack of Awareness	Lack of Emotion	Not Possible	Other	Total
Male	47	24	9	24	17	65	25	13	224
Female	53	19	5	13	10	32	7	9	148
Total	100	43	14	37	27	97	32	22	372

Table 4: Clustered contingency table for age and the Unamused face emoji. The Other category stands for Sadness, Surprise, Fear, Trust, Joy, Anticipation and Lack of Awareness

Age Range	Anger	Disgust	Lack of Emotion	Other	Total
18 - 23	36	13	8	10	67
24 -29	24	23	27	36	110
30 - 35	14	22	12	24	72
36 - 41	11	18	6	20	55
42+	8	17	9	45	79
Total	93	93	62	135	383

spelling mistakes or alter the responses in any way before labeling. We identify four response categories.

(1) One word in English, representing an emotion
(2) One word in English, not representing an emotion
(3) Multiple words or descriptive responses in English
(4) Non-English responses

A response in the **first category** was chosen as specified in Plutchik's wheel of emotions [23] because this model has eight basic emotions: Anger, Disgust, Sadness, Surprise, Fear, Trust, Joy and Anticipation. For the single-word responses that could not be directly mapped to the emotions from Plutchik's framework, we used *WordNet* and the website http://www.thesaurus.com to determine synonyms. The **second category** was a response that was clearly not an emotion such as "alien sneezing".

The **third category** contained multiple words or sentences, and sometimes one of the words in the response was a basic emotion. In this case, the response was categorized under that basic emotion category. To prevent oversimplification of the interpretation of the responses, we decided not to do the same if the response included a synonym or a related emotion. Also responses such as "angry or disgusted" were not categorized automatically: WordNet lemmatizers consider for example 'anger' and 'angry' to be different words as they are different parts of speech. Responses that could not be automatically mapped to emotions, were manually labeled by each of the authors taking an average of 5 hours per labeler. A detailed labeling manual created by the first author was used for the labeling task. The labelers were allowed to choose from 11 options; the eight basic emotions from Plutchik's wheel, "Lack of Awareness", "Lack of Emotion", and "Not Possible". The labelers were explicitly instructed, in the case multiple emotions were relevant, to choose the most relevant emotion, and only to resort to "Not Possible" as the last option. Due to moderate agreement, we chose to use a majority vote to set a label. The responses labeled differently by each

Table 5: Clustered contingency table for gender and the Nauseated Face emoji. The Other category stands for Anger, Sadness, Surprise, Trust, Joy, Anticipation and Lack of Awareness.

Gender	Disgust	Fear	Lack of Emotion	Not Possible	Other	Total
Male	112	5	74	15	18	224
Female	81	5	45	6	11	148
Total	**193**	**10**	**119**	**21**	**29**	**372**

labeler were discussed during a joint meeting that lasted 3 hours. The **fourth category** consisted of all non-English responses. All responses were labeled by native speakers of that language using the same detailed labeling manual used in the third category.

After labeling and categorization, we created two contingency tables for each emoji; one with labels and gender and the other with labels and age categories. Because we did not have enough data regarding participants who did not identify as male or female, we chose to exclude these responses. Some emojis had a clear and consistent interpretation, for these emojis no statistical analysis was performed. For the other emojis, we perform a chi-square (χ^2) test. The null hypothesis of χ^2 states that the interpretation of the emoji is independent on the category (gender or age, in our case). The alternative hypothesis states the opposite, i.e., the interpretation depends on the category. For some emojis, not all cells in the contingency table have a value larger than 5, we chose to combine the rows in which at least one of the cells is lower than 5. This is necessary to make the chi-square test work [1, 19].

5 STUDY RESULTS

Responses were given in English (3342), Dutch (57), Russian (50), French (8) and Portuguese (17). In total, we had 3474 responses for all emojis, of which 1380 were unique. After initial labeling, Kappa's agreement between the three labelers was moderate (0.468). After discussing the 180 responses for which no majority vote could be applied, an agreement was reached on all of them. The age distribution can be found in Table 2. From the 386 participants, 224 were male, 148 were female, 11 non binary and 3 responded as "Other".

We start by observing that across all gender and age categories, at least 94% of the respondents recognize the *Pouting Face* emoji 😡 as anger. This is by far the most agreed upon emoji in our collection. Similarly, clear agreement across all gender and age categories has been observed for such emojis as *Face with Tears of Joy* 😂 (joy), *Face with Open Mouth* 😮(surprise) and *Crying Face* 😢 (sadness).

Table 6: Clustered contingency table for age and the Face Screaming in Fear emoji.

Age Range	Surprise	Fear	Total
18 - 23	43	19	62
24 -29	61	41	102
30 - 35	34	31	65
36 - 41	21	31	52
42+	30	37	67
Total	**189**	**159**	**348**

Table 7: Most common response per category per emoji

Category	😡 (Anger)	😂 (Joy)	😯 (Surprise)	😢 (Sadness)	Total
Male	221	213	179	214	224
Female	144	143	127	141	148
18-23	66	66	61	65	67
24-29	109	104	97	105	110
30-35	72	69	66	68	72
36-41	54	53	48	53	55
42+	75	74	63	75	79

In Table 7 the results can be seen for each emoji per category. The text under the emoji is the most common interpretation of the emoji. Each cell gives the amount of responses that gave this response. The numbers under total represent the number of participants in that category, and are put there for easy reference.

Next, we observed that **the suitability of several emojis to probe for emotional state in gender- and age-diverse target groups vary per emoji**. The interpretation of the *Unamused Face* emoji 😒 differs per age category: younger people are more inclined to interpret this emoji as *anger*, whereas this shifts to *disgust* and further towards other interpretations (of which the biggest share comes from *sadness*) as age progresses (see also Table 4). This observation that the interpretation of the *Unamused Face* emoji depends on age is further supported by the χ^2 test ($p < 0.0001$). However, with respect to gender, we cannot reject the null hypothesis (($\chi^2 = 7.423$, $d = 5$, $p = 0.191$), hence we cannot say the interpretation of this emoji is influenced by gender. This observation concurs and reinforces the earlier observation of Miller et al. [18] that the *Unamused Face* emoji has the highest semantic misconstrual score among the Apple emojis studied.

Similarly, in case of the *Face Screaming in Fear* emoji 😱, the interpretation shifts from *surprise* to *fear* from younger to older age groups, see Table 6. This indicates that the *Unamused Face* and *Face Screaming in Fear* emoji can only be used if all participants are in the same age category, but results might be incorrect if young and older people are mixed within the study. Hence, one needs to be cautious when using *Unamused Face* and *Face Screaming in Fear* emojis for emotional awareness.

The *Face With Rolling Eyes* emoji 🙄 is age independent, but gender dependent; here females indicate the emoji as *anger*, whereas males interpret it more neutrally. The results for this emoji can be seen in Table 3. Hence using this emoji in a study containing both genders might give incorrect results as well. Since there is a clear emoji for *anger* (See Figure 1), we would advise not to use the rolling eyes emoji, especially not in gender-diverse groups.

Table 8: Summary of the emojis' interpretation

Emoji	Emotional State Interpretation
😡	Anger
😒	Anger, Disgust, Sadness, Lack of Emotion
🤢	Disgust, Lack of Emotion
😱	Fear, Surprise
🙄	Anger, Disgust, Surprise, Lack of Emotion
😢	Sadness
😂	Joy
😯	Surprise
😐	Lack of Emotion

Finally, the *Nauseated Face* 🤢 and *Neutral Face* 😐 emojis are interpreted independently by gender and age. For the *Nauseated Face* emoji, although there is a small favor towards *digust*, there is no clear winner in all cases. Quite often it is seen as an object not representing emotion but, e.g., physical condition such as vomiting. When analyzing the contingency tables, for the *Nauseated Face* emoji, there is a majority representing *disgust*, which is in accordance with one of the definitions from Emojipedia. However, there is a close runner up, namely *Lack of Emotion*. See Table 5 for a clustered contingency table for the *Nauseated Face* emoji. This is in contradiction with its definition on Emojipedia. It would thus not be useful to include this emoji when probing for one's emotional state.

The *Neutral Face* emoji should represent a neutral state, and although the emoji is also often interpreted as *surprise* (especially by males), *lack of emotion* is the most often used interpretation. The *Neutral Face* emoji can be used for finding an emotional baseline.

Main Finding: Basic emotions of anger, surprise, joy and sadness had conclusive results and were understood as intended by their corresponding emoji. However, disgust and fear (for younger age groups) did not have conclusive results. Younger people are more inclined to interpret the *Unamused Face* emoji as anger.

6 THREATS TO VALIDITY

It is possible that people's use of emojis is highly dependent on the context it is used in. However our goal was to situate this in a more general context of emotion awareness and to do this, only emojis related to basic emotions were used. It is also possible that other factors besides age and gender such as culture, disability, socioeconomic status, occupation and the context in which the emoji is used (to name a few) could affect the results. However, we

wanted to start by looking at age and gender in particular and later expand to include other factors in a larger followup study.

7 CONCLUSIONS AND FUTURE WORK

The paper presents the results of an online survey seeking to determine the validity of using emojis for emotional awareness of the participant in a research study. The results show that six basic emotions of anger, disgust, fear, surprise, joy and sadness had conclusive results and were understood as intended. However, disgust and fear (for younger people) did not have conclusive results. To answer our research question, we can thus conclude that there are emojis (Unamused Face, Nauseated Face and Face With Rolling Eyes) that are being interpreted differently depending on age or gender. However, there is a collection of emojis (Pouting Face, Crying Face, Face with Tears of Joy, Face with Open Mouth, Neutral Face) which are indeed suited for determining emotional awareness in gender and age diverse target groups. Our goal was to situate in a general context of emotion awareness. As we needed to start somewhere, we only considered gender and age. However, other factors such as context and demographic and cultural factors also play a role in the interpretation of an emoji. These threats to validity should be considered in future work. Other setups for research can be thought if as well. For example, the participant must choose between several emojis, or must label text sentimentally with emojis.

As part of our future work, we are interested in learning how software developers use emojis, and to see whether it can then be inferred from emoji use what the emotional state of the software developer is. Emojis can also help improve the working environment, improve user friendliness and help the developer to become emotionally aware of his/her own emotions. Besides more research needed on culture and other emotional models with other labelers, as some persons who did not identify as male or female participated in the survey. These results might be different. Besides this, further research can include the interpretations of other emojis to see whether different ones or different visualizations of emojis obtain better results. This research presents a systematic analysis which will be useful to research similar kinds of questions on this topic.

ACKNOWLEDGMENTS

This work is supported in part by grants from the National Science Foundation under grant numbers CCF 18-55756 and CCF 15-53573. We are grateful to Alexandra Fountaine, who provided support and ideas for the survey. We thank all volunteers who participated in our survey. We are also thankful to Stephan van den Berkmortel, Weslley Silva Torres and Nathan Pottier for labeling non-English responses.

REFERENCES

[1] Alan Agresti and Christine A Franklin. 2007. Statistics: The Art and Science of Learning from Data (3rd). (2007).
[2] Zhenpeng Chen, Xuan Lu, Sheng Shen, Wei Ai, Xuanzhe Liu, and Qiaozhu Mei. 2017. Through a Gender Lens: An Empirical Study of Emoji Usage over Large-Scale Android Users. *arXiv preprint arXiv:1705.05546* (2017).
[3] Nicholas Clarke. 2010. Projects are emotional: How project managers' emotional awareness can influence decisions and behaviours in projects. *International Journal of Managing Projects in Business* 3, 4 (2010), 604–624.
[4] James A Coan and John JB Allen. 2007. *Handbook of emotion elicitation and assessment*. Oxford university press.
[5] Pauline Dewan. 2015. Words Versus Pictures: Leveraging the Research on Visual Communication. *Partnership: the Canadian Journal of Library and Information Practice and Research* 10, 1 (2015).
[6] William Foddy and William H Foddy. 1994. *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge university press.
[7] Octavio García, Jesús Favela, and Roberto Machorro. 1999. Emotional awareness in collaborative systems. In *String Processing and Information Retrieval Symposium, 1999 and International Workshop on Groupware*. IEEE, 296–303.
[8] Daniela Girardi, Filippo Lanubile, Nicole Novielli, Luigi Quaranta, and Alexander Serebrenik. 2019. Towards Recognizing the Emotions of Developers Using Biometrics: The Design of a Field Study. In *International Workshop on Emotion Awareness in Software Engineering*.
[9] Judith A Hall and David Matsumoto. 2004. Gender differences in judgments of multiple emotions from facial expressions. *Emotion* 4, 2 (2004), 201.
[10] Michael B Hudson, Sylis C Nicolas, Molly E Howser, Kristen E Lipsett, Ian W Robinson, Laura J Pope, Abigail F Hobby, and Denise R Friedman. 2015. Examining how gender and emoticons influence Facebook jealousy. *Cyberpsychology, Behavior, and Social Networking* 18, 2 (2015), 87–92.
[11] Sara R Jaeger, Yixun Xia, Pui-Yee Lee, Denise C Hunter, Michelle K Beresford, and Gastón Ares. 2018. Emoji questionnaires can be used with a range of population segments: Findings relating to age, gender and frequency of emoji/emoticon use. *Food Quality and Preference* 68 (2018), 397–410.
[12] Eva Jonas, Verena Graupmann, and Dieter Frey. 2006. The influence of mood on the search for supporting versus conflicting information: Dissonance reduction as a means of mood regulation? *Personality and Social Psychology Bulletin* 32, 1 (2006), 3–15.
[13] Robbert Jongeling, Subhajit Datta, and Alexander Serebrenik. 2015. Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 531–535.
[14] Robbert Jongeling, Proshanta Sarkar, Subhajit Datta, and Alexander Serebrenik. 2017. On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering* 22, 5 (2017), 2543–2584.
[15] James Lani. [n. d.]. Profile of Mood States (POMS). ([n. d.]).
[16] Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*. Association for Computational Linguistics, 63–70.
[17] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 770–780.
[18] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. *"Blissfully happy" or "ready to fight": Varying interpretations of emoji*. AAAI press, 259–268.
[19] Douglas C Montgomery and George C Runger. 2011. *Applied statistics and probability for engineers (5th)*. John Wiley & Sons.
[20] Sebastian C Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: sensing developers' emotions and progress. In *ICSE*. IEEE, 688–699.
[21] Yukiko Nishimura. 2016. A sociolinguistic analysis of emoticon usage in Japanese blogs: Variation by age, gender, and topic. *AoIR Selected Papers of Internet Research* 5 (2016).
[22] Allan Paivio and Kalman Csapo. 1973. Picture superiority in free recall: Imagery or dual coding? *Cognitive psychology* 5, 2 (1973), 176–206.
[23] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.
[24] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*. 1631–1642.
[25] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai. 2010. Sentiment in short strength detection informal text. *JASIST* 61, 12 (2010), 2544–2558.
[26] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.
[27] Alecia Wolf. 2000. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior* 3, 5 (2000), 827–833.
[28] Michal R Wrobel. 2013. Emotions in the software development process. In *HSI*. IEEE, 518–523.

When a Merperson is a Merman

Using Gender Inclusive Designs for Codepoints which Do Not Specify Gender

Jennifer Daniel
Google Inc.
San Francisco, CA, USA
jenniferx@google.com

ABSTRACT

Currently, to support gender inclusive codepoints (for example, gender inclusive can be defined as male/female to an equal degree, can neither be confidently identified as male/female, etc.) all major platforms default to a male or a female design. So, if someone were to send a text to their friend from a Microsoft device, **"Love a good mansplain** 🤦‍♀️**"** their friend, if reading from an iPhone will see, **"Love a good mansplain** 🤦‍♂️**"** even though both of these emojis map to U+1F926. This creates all kinds of cross platform inconsistencies and in some cases reinforces stereotypes.[1]

CCS CONCEPTS

• **Applied computing** → **Media arts**.

KEYWORDS

Gender; Emoji; Gender Inclusivity; Representation; Abstraction and Realism; Emoji Design; Unicode

ACM Reference Format:
Jennifer Daniel. 2019. When a Merperson is a Merman: Using Gender Inclusive Designs for Codepoints which Do Not Specify Gender. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316542

1 INTRODUCTION

"Other than the above list (see Figure 3), human-form emoji should normally be depicted in a gender-neutral way unless gender appearance is explicitly specified" – *TR-51, Emoji Documentation* [2].

As of Unicode 12.0, 64 emojis are defined as "gender inclusive" yet most platform renderings do not have unique designs to support the vast majority of these codepoints which do not specify gender [1] (While it varies from vendor to vendor you could reasonably say 53 emojis do not have unique gender inclusive designs.). The only explicitly gender inclusive emojis all platforms support with unique designs are for Child 🧒 U+1F9D2, Person 🧑 U+1F9D1, Older Person 🧓 U+1F9D3[2].

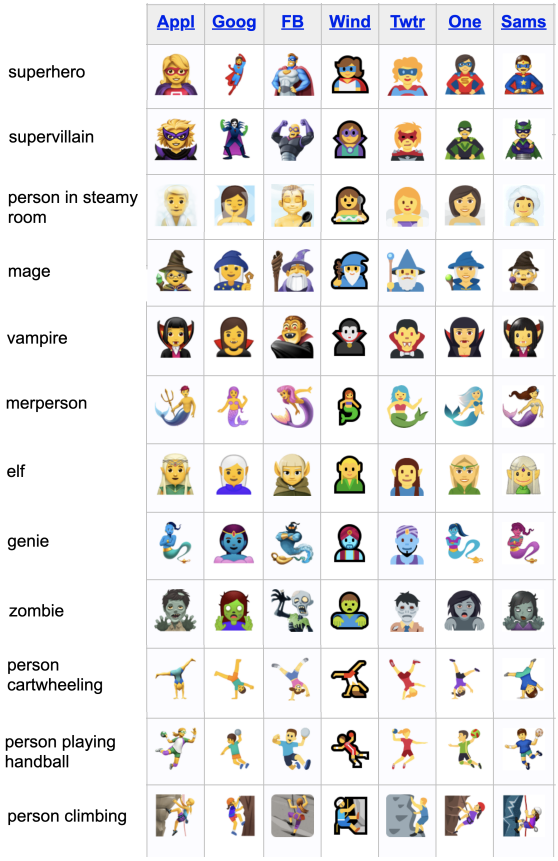


Figure 1: "Gender Inclusive" codepoints across platforms: Sometimes, a merperson is a merman. What gender are you once you're dead? or the living dead? (Image Source: Composite by Jennifer Daniel/Unicode.org).

Giving "ungendered" emojis a gender inclusive appearance is not trivial. Are there signifiers that can provide clues which don't rely on a gender binary (haircut, clothing, color, body language, other

[1]Example: "Construction worker" 👷 (U+1F46E), "Police officer" 👮 (U+1F477) etc. present as male.

[2]Outliers include Google, the only platform to support designs for gender inclusive Kiss 💏 (U+1F48F), Family 👪 (U+1F46A), Couple with heart 💑 (U+1F491), and, Samsung, the only platform to support gender inclusive designs for superhero 🦸 (U+1F9B8) and Supervillain 🦹 (U+1F9B9).

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316542

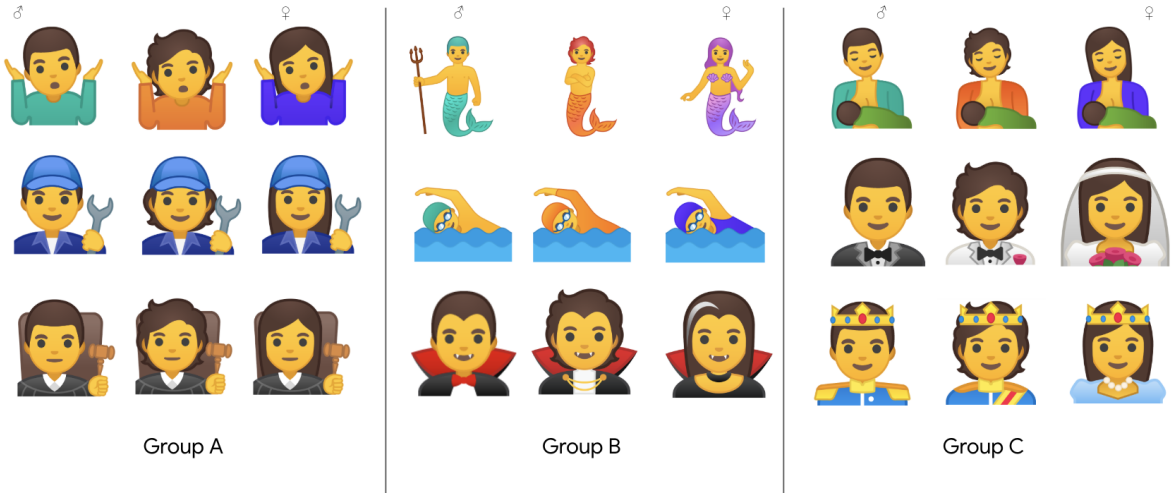


Figure 2: Male, Gender inclusive, and Female emojis: Some emojis could simply require a change in hairstyle or clothing color (Group A); Others require more refinements in accessories, props and body language (Group B); And others may force us to look each other in the eye and think deeply about gender (Group C).

Female		Male	
U+1F467	girl	U+1F466	boy
U+1F469	woman	U+1F468	man
U+1F475	old woman OLDER WOMAN	U+1F474	old man OLDER MAN
U+1F46D	women holding hands TWO WOMEN HOLDING HANDS	U+1F46C	men holding hands TWO MEN HOLDING HANDS
U+1F46B	woman and man holding hands MAN AND WOMAN HOLDING HANDS		
U+1F936	Mrs. Claus MOTHER CHRISTMAS	U+1F385	Santa Claus FATHER CHRISTMAS
U+1F478	princess	U+1F934	prince
U+1F483	woman dancing DANCER	U+1F57A	man dancing
U+1F470	bride with veil		
U+1F930	pregnant woman		

Figure 3: Emoji With Explicit Gender Appearance

facial features)?[3] What level of detail is too much or not enough? Emojis demand an instant read so what markers will be effective at communicating a spectrum of gender presentations at a small size?

Gender is a construct. When exploring signifiers that have potential to communicate if a design is inclusive it is important to remember that gender lives dynamically on a spectrum and there is no single visual design "solution". However, we cannot ignore gender and Unicode continues to support existing and possibly future codepoints.[4] Gender inclusive designs are intended to represent a person of any gender. You could argue that a gender inclusive design 'works' if the gender inclusive artwork is truly ambiguous to the viewer. But, due to each individual's own perception or ascribing of gender, it's possible people look at the existing "male" emoji designs and see "woman with short hair" or look at the "female" emoji designs and see "man with long hair". Given how fluid and dynamic language and gender presentation can be – it's not inconceivable that regardless of haircut, shirt color, accessories, and other clues, gender inclusive emojis could be repurposed in some other manner and bring us somewhere else entirely.

2 ACKNOWLEDGEMENTS

I would like to thank Shiraz Fuman, Paul Hunt, Seb Grub, Jeremy Burge, Ben Hutchinson, and Sam Jaffee for their guidance and support in this effort.

REFERENCES

[1] Charlotte Buff. 2017. Proposal for Fully Gender-Inclusive Emoji. (2017).
[2] Mark Davis and Peter Edburg. 2019. ESC. Unicode® Technical Standard #51. (2019). http://unicode.org/reports/tr51/

[3] Paul D. Hunt (2016). Proposal to enable gender inclusive emoji representation, Unicode. https://www.unicode.org/L2/L2016/16317-gender-inclusive-emoji.pdf
[4] Mark Davis, Jeremy Burge, ESC (2018). Gender Emoji Strategy. https://www.unicode.org/L2/L2018/18022-emoji-gender.pdf

Emoji Sequence Use in Enacting Personal Identity

Jing Ge

Department of Anthropology, University of California, Berkeley, California, U.S., jingge@berkeley.edu

ABSTRACT

Focusing on a Chinese social media platform, this study adopts computer-mediated discourse analysis to examine how users employ emoji sequences to construct their personal identity through the expression of stance and engagement. Seven types of linguistic elements were identified by conducting stance and engagement analysis on emoji sequences in posts by social media influencers. Stance was more frequent than engagement. Attitude markers were the most common element used to convey stance, whereas directive was the most prevalent element used to express engagement. In addition, emoji sequences that did not convey stance and engagement were coded as n/a. This study also observed creative usages in the composition of emoji sequences that compensate for the lack of a prescribed emoji sequence grammar. Based on these findings, it advances recommendations for the design of emoji and of social media platforms grounded in linguistic principles.

CCS CONCEPTS

CCS → Applied computing → Law, social and behavioral sciences → Anthropology

KEYWORDS

Emoji sequences, Language, Identity, Self-presentation, Stance and engagement

1 Introduction

The common occurrence of emoji sequences in computer-mediated communication (CMC), defined as "two or more emoji with different meanings that stand together to form a conceptual unit" [12, p.1], has led to the popular speculation that emoji are evolving into a new and digitally-mediated language [8, 12]. Indeed, these colorful, cute and playful pictographs go beyond merely expressing emotions and embellishing messages; rather, a sequence of emoji can function as a sentence-like textual utterance (though it may not convey a complex idea or concept), offering social media users a new way of meaning-making, expression and self-presentation [8, 15, 26]. Well-known instances described in the popular press include Roger Federer using an emoji-only tweet to describe his day off and Andy Murray employing a string of emoji to announce his wedding [13]. Previous CMC research on self-presentation and identity construction also provide some examples that illustrate emoji sequence usages [26], according to the above-mentioned definition.

The emergent practices of selecting and grouping emoji together for self-narrative raise challenges for theoretical and practical understanding of emoji use in CMC. Emoji are created primarily by professionals, and their adoption is determined by different social media platforms. Thus, the design of emoji and of social media platforms need to be grounded in an understanding of the innovative and complex ways users employ emoji to engage in social interactions. Moreover, people construct and negotiate their identity each time when they communicate using a new language [17]. If emoji are to develop as an independent graphical language, users need to have emoji not only expressing emotions but also constantly organizing and re-organizing who they are and how they relate to the digital world. Existing emoji literature only mentions the role of single emoji in the establishment of user identity [7, 24, 36, 46, 50]. Further, the few previous studies that pertain to emoji sequences mainly focus on their pragmatic and communicative functions [12]; their social use, such as when employed to enact personal identity, remains largely unexplored.

Based on the premise that emoji sequences are functionally analogous to sequences of words in verbal language, this study explores how social media users employ such sequences to enact their personal identity

by using computer-mediated discourse analysis [14], an approach that adapts methods from linguistics and related fields to analyze digital discourse. Specifically, it employs stance and engagement analysis to investigate how users select and use emoji as linguistic elements to express their 'textual voice' and community recognized personality and also to encourage their audiences for being discourse participants [20]. Sina Weibo, China's indigenous microblogging site, lends itself well to emoji sequence analysis. Emoji use – including the formulation of emoji sequences – is common on Weibo due to the platform's emoji affordances and developed emoji culture [11, 12]. In short, by focusing on China, whose culture fosters an especially high concentration of emoji use [11], this study advances current CMC and emoji literatures, while also providing new insights which can inform the future design of emoji and social media platforms.

In what follows, this paper first develops this theme in relation to personal identity, self-presentation on social media, and emoji sequence literatures, after which it provides background information on stance and engagement analysis. It then describes its dataset, which consists of Sina Weibo posts published by Chinese celebrities, and also the analytical procedures followed. Overall, this study found that 95% of the emoji sequences were analyzable as conveying linguistic elements required for the expression of stance and engagement, supporting the premise that these textual utterances can be used for enabling one's personal identity. Further, stance was found to be more prevalent than engagement. 'Attitude markers' (i.e. indicating a message author's affective attitude to propositions) is the most common linguistic element for expressing stance, and 'directive' (i.e. asking readers to perform an action or show their agreement) is the most common element for conveying engagement. This study also observed creative use of emoji repetition to compensate for the lack of emoji which can function as adverbs, thereby making certain sequences more 'language-like.' Based on these findings, this study advances recommendations for emoji design and social media platforms grounded in linguistic principles.

2 Background

2.1 Operating the concept of self-identity

The common consensus is that "language use is a form of self-presentation" – people not only convey and exchange information with their interlocutors but attach their own identity as well as their relation to the social world to "particular written and spoken texts" [28, p.293]. The term 'personal identity' refers to a set of individuals' attributes, beliefs, and desires which allow them to distinguish themselves in socially relevant ways [10]. People attempt to assert their individuality but also try to connect with others to meet their desire for recognition, affiliation, and a sense of belonging [30]. In seeking to portray their identity, such individuals implement 'identity-work' through their choices about how to verbally communicate [49]. Through considering emoji as 'words', this study examines personal identity from the perspective of how social media users employ emoji sequences as a way of self-presentation.

2.2 Self-presentation on social media

Drawing on the social dimension of social media, the concept of self-presentation states that individuals engage in a process of impression management and information control while interacting with both actual and imagined audiences in networked digital space [24]. Specifically, people use social media-afforded multimodal communication (such as text, images and website links) to convey their opinions, emotional states and intentions [3] and to construct the discourse context in which their identity can be interpreted and understood by others at the same time [34, 47].

While some scholars have focused on textual self-prestation [e.g., 4, 9, 17, 48], others have studied visual impression management through examining profile photos and selfies [e.g., 5, 23, 37, 41, 43]. For instance, [9] investigated how online dating participants use well-crafted textual messages to present an ideal self. In examining profile statuses of WhatsApp Messenger, [17] found that users employ a variety of linguistic elements (e.g. 'self-mention' 'directives') to not only express their stance but acknowledge their audiences

as discourse participants. In contrast, [41] compared online visual self-presentation from a gender perspective, revealing that girls tend to focus on both the aesthetic, emotional and symbolic aspects of photographs more than their male counterparts. Similarly, [23] also identified such gender differences with regard to gaze, posture, dress, and distance from the camera, and concluded that the teens construe their profile images as invitations to interact with others online. Moreover, some of the previous studies focused on a combination of visual and textual self-presentation [e.g. 16, 40, 42]. The preceding review suggests the CMC literature primarily focuses on the topic of enacting personal identity through text and photographs. The emergent digitally-mediated graphical language, that is emoji, remains unheeded.

2.3 Emoji sequences

Previous studies primarily focus on the pragmatic and communicative functions of emoji sequences. This line of research sheds light on users' tendency to group different emoji together to convey complete messages [8, 15, 12]. [15] identified 'narrative sequence' (i.e. the use of a series of consecutive emoji to tell a story) as one of the pragmatic functions of emoji. 'Narrative sequence' aligns with 'emoji phrase' or the use of a string of emoji in the expression of propositions [8] Moreover, [12] showed that emoji sequences communicate a variety of speech acts (such as making a claim and expressing desire), and exhibit different rhetorical relations with the accompanying texts (such as restatement and elaboration).

There are a few studies that, while not focusing directly on emoji sequences, do include examples that illustrate how users employ sequences to make meaning and also verbally interact with others in CMC. In one study of the pragmatic intentions underlying emoji use [6], several users reported that using instances of emoji sequences provides speech context in which their messages can be correctly interpreted and understood by their peers. [11] examined emoji use for persuasion, and observed that users sometimes employ two or more different emoji to express emotions (i.e. pathos), to provide detailed information (i.e. logos), and to create similarity with message readers (i.e. ethos). Instances derived from research on emoji hashtags showed that Twitter and Instagram users tag a cluster of emoji to convey a message subject [18] and to express emotional statement [32].

By considering emoji as 'words', a number of studies observed grammatical (or at least grammar-like) properties within emoji sequences [29, 35, 39, 12]. In exploring the translation of text to 'emoji sentences', [29, 35] found that emoji that represent facial expressions and actions/activities often function as verbs while emoji that denotate objects function as nouns. These findings partially align with [12]: The authors observed that, in order to employ emoji sequences to convey speech acts, Weibo users not only use emoji that represent facial expressions and body gestures as explicit verbs but also emoji that denotate objects as implicit verbs; users also tend to use animal emoji that indicate certain personalities and attitudes as well as gender symbols as pronouns. In contrast, [38] and [39] noticed that YouTube and WhatsApp users often string two or more identical emoji together as an 'intensifier' to modify their assertions, which is similar to how an adverb functions. Relatedly, [27] have argued that repeated emoji are similar to 'beat gestures' in speech. Other scholars have commented on the syntactic relations within sequences, to some extent. For instance, [8, 45] showed that emoji sequences exhibit subject- or stance-first word order. (This is also normal word order in Chinese.) [8] also revealed an iconic conceptual structure in some strings of emoji, as well as the practice of calquing, whereby emoji are directly mapped onto morphemes, words, or utterances of verbal language. Conversely, [12, 44] observed that Weibo and Google Hangout users tend to place stance-indicating emoji after emoji that represent actions, activities and objects (similar to verb-subject and object-verb word order).

In sharp contrast, among previous studies only a few but merely acknowledge the role of single emoji in the construction of personal identity [7, 15, 34, 36, 46]. For instance, [46] reported that, among Japanese users, using emoji in an aesthetic manner is an important aspect of expressing and managing one's online self-image. [36] noted the practice of Twitter users using various emoji skin tones to present their ethnicity. In

examining emotional self-presentation on WhatsApp, [26] included some examples that illustrate how users sometimes employ a cluster of emotion-laden emoji to express their attitudes, opinions and feelings.

3 Methodology

A review of the literature shows that language use on social media is a form of self-presentation; emoji sequences can substitute for sequences of words and are functionally analogous to verbal language. Thus, a working premise of this study is that emoji sequences potentially express verbal meaning (although it may not always be straightforward to translate them into words), constitute language properties, and allow for self-presentation. To explore this premise systematically, this study employs a language-focused approach – stance and engagement analysis [20] – to address the following research question:

RQ: Which linguistic elements are formulated in emoji sequences by users to enact their personal identity?

3.1 Stance and engagement

The literature on emoji sequences suggests that they function as textual utterances in CMC [8, 15, 12]. These, in turn, allow users to construct their personal identity through the expression of stance and engagement [20]. Stance is considered as attitudinal dimension: it constitutes the ways writers express a textual 'voice' or community recognized personality, and convey their judgments, opinions, and commitments. Specifically, linguistic elements used to communicate stance include: 'hedges' (e.g. possible, might), 'boosters' (e.g. highly, very), 'attitude markers' and 'self-mention' (e.g. I, our). Furthermore, engagement is seen as an alignment dimension: it encompasses the ways writers acknowledge the presence of their audiences and encourage them for discourse participation. Linguistic elements used to convey engagement include: 'reader mention'(e.g. you, your), 'directives' (e.g. come, go) 'questions', 'appeals to shared knowledge' (e.g. of course, be well recognized) , and 'personal asides' (e.g. as mentioned).

Noteworthy is that the primary focus of stance and engagement analysis has been on written discourse, especially academic writing [e.g., 21], but it has also been applied to spoken and multimedia discourse [e.g., 17, 22]. For instance, it has been adopted by [17] to analyze how WhatsApp users employ their profile statuses to reveal their personal identity. The advantages of using stance and engagement analysis are that they are well-defined and have been tested extensively for textual materials [21]; this is the first study to apply this approach to an analysis of emoji.

3.2 Data

The data comprise two types of Sina Weibo posts: 1) combinations of emoji sequences and text; 2) independent emoji sequences. All of these are operationalized as instances of two or more semantically different emoji grouped together with no intervening text or punctuation. Because emoji sequence use in enacting personal identity is a new line of inquiry, this study sought information-rich samples that captured the phenomena of greatest interest [1]. This led this study to sample emoji sequences as used by social media influencers (celebrities or entertainment professionals, including actors/actresses, singers, and TV show hosts). It is well recognized that such figures are adept at using emoji sequences (as defined by this study) to convey playful and persuasive messages [11, 12]. The sampled influencers thus are particularly important for informing and advancing the understanding of emoji sequence use.

This study used the 'Sina Weibo Influencer Popularity List' (http://data.weibo.com/top/hot/famous), selecting 87 accounts owned by 87 celebrities, based on the criterion that the most recent 10 posts had at least one post containing an emoji sequence. It then manually collected 403 Weibo posts that contain emoji sequences posted between April 11 and April 28, 2017 and between November 11 and November 28, 2018. This size is considered effective for a manual coding study [25].

3.3 Analysis method

This study adopted computer-mediated discourse analysis, at the level of text analysis, [14], to analyze and identity the linguistic elements involved in the expression of stance and engagement. To this end, it was first necessary to translate them approximately into Chinese words. This study conducted three-step emoji sequence translation in an attempt to overcome its subjective tendency. First, the author, a native Chinese speaker and an emoji researcher, translated and coded the entire data set. The author also coded a large amount of the data together with: 1) an English-speaking emoji researcher who understands the linguistic structure of Chinese; and 2) a Chinese scholar whose expertise is on cultural communication. The emoji definitions provided by Weibo helped during this process. At the end of the first step 79 ambiguous sequences remained. This study then conducted a second step by recruiting nine crowdsourced coders on Witmart (http://www.witmart.com/cn/, a Chinese crowdsourcing platform that is equivalent to Amazon's Mechanical Turk) to provide interpretations of the ambiguous sequences. For the remaining 20 cases of unclear examples, in a third step the author consulted three emoji consultants recommended by a Chinese graphicon design firm. All 20 ambiguous cases were solved.

All of the emoji sequences in the Weibo data were coded in the discourse context rather than in isolation. Specifically, the repetitive use of identical emoji for intensification is functionally equivalent to the use of adverbs [38, 39], and was coded as 'boosters'. Emoji that convey sentiments and opinions (e.g. hearts, facial expressions, body gestures) were coded as 'attitude markers'. This study coded 'self-mention' and 'reader mention' by identifying emoji that function as first- and second person pronouns, depending on specific context. For example, animal emoji often function as 'I' and 'you'; gender symbols represent 'I'; emoji that denotate 'onlookers' serve as 'we' and 'you'. Moreover, emoji that show 'a winking face' and 'a winking face showing a stuck-out tongue' allow for the conveying of shared secrets [8], and were coded as 'appeals to shared knowledge'.

Emoji functioning as deictic and imperative verb were coded as 'directives', such as 🫳 (meaning 'come') and 🫂 (defined by Weibo as 'follow each other'). Emoji that denotate question mark and represent puzzled facial expressions and body gestures were coded as 'questions'. Those examples which could not be coded by applying stance and engagement analysis were coded as 'n/a', including sequences comprising a list of emoji functioning as nouns. The results of the coding are provided in the next section.

4 Findings

The mean number of emoji used across all sequences is 5 with a mode of 4. The Standard Deviation (SD) is 3.52. The scope of emoji represented in the data includes not only all the categories in the Emoji Unicode 12.0 list (e.g. smileys and people, animals and nature, objects, activities, food and drink, travel and places, symbols, flags) but also specific Sina Weibo-afforded ones such as animated emoji and emoji including text overlays. In what follows, the frequency distribution results of the stance and engagement analysis are presented first, followed by examples illustrating each category.

4.1 Stance and engagement

Analysis of the Weibo data shows in order to enact personal identity, the users prefer to use emoji sequences for the expression of stance more than of engagement. As Table 1 presents, 'attitude markers' accounts for the largest overall percentage, followed by 'self-mention' and 'boosters'. In their attempts to comment on specific people and/or issues mentioned in the sequences or text, the social media influencers in the dataset often used sentiment-laden emoji (e.g. emoji that represent heart(s), facial expressions and body gestures) to convey their judgments, opinions and emotions. Emoji that convey 'attitude markers' include emoji that function as verbs (example 1, 3) and as adjectives (example 4) in the discourse context. Specifically, emoji that function as adjective comprise two general categories: the first can be directly translated into adjectives based on the definition offered by Weibo (e.g. 😊, defined as 'lovely'); the second is derived from applying

the rebus principle. For example, 💪 which means '威武' (or 'great/awesome') comprises the letter 'V' and a Arabic number '5'. The pronunciation of 'V' in English and of '威' in Chinese is identical; the pronunciation of '5' and '武' in Chinese is also identical. That is, a combination of 'V' [wēi] and '5' [wǔ] means '威武' [wēi] [wǔ].

Table 1: Stance and engagement: Linguistic elements and frequency distribution

	Frequency	Total	Percentage
Stance			
Attitude markers	255	403	63.3%
Self-mention	123	403	30.5%
Boosters	121	403	30%
Engagement			
Directives	70	403	17.4%
Reader mention	37	403	9.2%
Questions	24	403	6%
Appeals to shared knowledge	11	403	2.7%
n/a	20	403	5%

The sample users sometimes employed emoji that function as first-person pronoun to communicate 'self-mention' in specific speech contexts. Specifically, to refer to themself ('I'), they tended to deploy gender symbols and those animal emoji that have certain personalities and express certain attitudes (example 2); some of the sample users (i.e. actor/actress) occasionally used their movie or TV names (example 1). The emoji that denotate multiple people (e.g. 'onlookers') were used as 'we' (example 4). while the 'booster' element was conveyed through emoji repetition. Weibo users in the dataset often grouped two or more emoji that have either identical appearance or similar semantic meaning to modify or intensify their assertions.

The engagement analysis revealed that 'directives' is the most prominent linguistic element used by the Weibo users in the data, followed by 'reader mention', 'questions', and 'appeals to shared knowledge'. All of the users preferred to employ emoji that function as deictic and imperative verbs to ask their message readers to do something, such as attend a concert (example 6) and follow one's Weibo updates (example 7). Similarly to 'self- mention', the animal emoji and emoji that denotate multiple people were used as second-person pronoun to refer to the message reader (s) ('you') (example 5). The use of 'question' for the expression of engagement is primarily conveyed through emoji that function as interrogative words (i.e., 'what' 'why'), including explicit and implicate ones (example 5). It was relatively infrequent in Weibo posts. Conversely, 'appeals to shared knowledge' is little-utilized linguistic element. The users in the dataset rarely used emoji that denotate 'a winking face' and 'a winking face showing a stuck-out tongue' as explicit markers when asking message readers to recognize something as familiar or accepted (that is, to convey something like 'you know I am kidding' or 'you know I mean') (example 7). Emoji sequences in the 'n/a' category are compositions of emoji that function as noun and therefore do not convey stance and engagement, as illustrated in examples. These examples are discussed in the following section.

4.2 Examples of emoji sequences

An emoji sequence may express engagement and/or stance (if the sequence includes the linguistic elements such as ‘self-mention’ and ‘question’). The following examples illustrate the identified linguistic elements employed by the Weibo users to enact their personal identity in the specific discourse context.

Example 1. An emoji sequence illustrating the expression of stance through use of an ‘attitude marker’ and ‘self-mention’.

终于尝到了传说中珍珠奶茶

A free English translation of the Chinese text in example 1 is: ‘(I) finally tried the popular Bubble tea.’ The first emoji in the sequence represents ‘Bubble tea’. The second emoji (defined by Weibo as ‘disappointment’) indicates the message author’s attitude towards this ‘Bubble tea’ and therefore serves as an ‘attitude marker’. The last animal emoji (‘fox’) is well recognized as the name the user’s name in a TV show, and thus functions as first person pronoun (‘I’) and serves as ‘self-mention’. The overall meaning is something like ‘Bubble tea, disappointment, I (say).’

Example 2. An emoji sequence illustrating the expression of stance through use of ‘self-mention’.

@ user

A brief background for example 2: The user posted this independent emoji sequence along with an image (a birthday gift given by his friend) below the sequence. The first emoji (i.e. making a bow with hands folded in front) means ‘thanks’; the last emoji indicates the user himself or ‘I’ [11] thus serves as ‘self mention’. The overall meaning is ‘thanks, I’.

Example 3. An emoji sequence illustrating the expression of stance through use of ‘attitude marker’ ‘booster’, and ‘self-mention’.

冬天来了，春天不在遥远了

A free translation of the text in example 3 is: ‘Winter has arrived. Spring is not far away’. The first emoji means ‘cheer on’ and shows the message author’s commitment, thus serving as an ‘attitude marker’. The following two emoji were used to emphasize his commitment and therefore function as ‘booster’. The last emoji (female gender sign) indicates the message author himself (‘I’) and therefore serves as ‘self-mention’.

Example 4. An emoji sequence illustrating the expression of stance through use of ‘attitude marker’, ‘booster’ and ‘self-mention’.

偶和偶的小伙伴们一致认为这才是真正的芭蕾

A loose translation of the text in example 2 is: ‘I and my friends think this is real ballet.’ The first two emoji denotate ‘man and women dancing’ mean ‘ballet’. The third emoji 💯 (meaning ‘good’) indicates the message author’s opinion and therefore conveys ‘attitude marker’; the following two emoji 💯💯 are used for intensification (i.e. ‘really’) [38, 39] and thus serve as ‘booster’. The last emoji (defined by Weibo as onlookers) means ‘we’ in this speech context and therefore functions as ‘self-mention’. The overall meaning of the sequence is ‘ballet, really good, we (think).’

Example 5. An emoji sequence illustrating the expression of engagement through use of ‘reader mention’ and ‘question’.

圣诞节 大家都想要些什么礼物呢

The Chinese text in example 3 means, ‘What Christmas gifts do you want to have?’ The first three emoji in the sequence represent ‘gifts’. The second (defined by Weibo as ‘question’) means ‘what’ and thus functions as ‘question’; the last two emoji (defined by Weibo as onlookers) refer to the message readers (‘you’) and thus convey ‘reader mention’. The overall meaning is ‘What gifts (do) you (want)?’

Example 6. An emoji sequence illustrating the expression of engagement through use of ‘directive’.

A brief background of example 4: The singer posted this independent emoji sequence along with an image (a poster advertising his concert) below the sequence. The first emoji (defined by Weibo as ‘come’) was used by the singer to request that message readers come to the concert and therefore serves as ‘directive’. The last three emoji that denotate music notes and microphone are positioned together to represent a conceptual idea, that is, ‘music/concert’. The overall meaning is something like ‘Come to (my) concert’.

Example 7. An emoji sequence illustrating the expression of engagement through use of ‘directive’ and ‘shared knowledge’.

关注我的动态了吗

A free translation of the text in example 5 is: ‘Have you paid attention to my news?’ The background is that this celebrity had announced her pregnancy on Weibo, generating a huge amount of comments and reposts. The overlay text in the first emoji (meaning ‘follow each other’) functions as ‘imperative verb’ and thus serves as ‘directive’. This celebrity used the second emoji (defined by Weibo as winking) to indicate the well-recognized news about her pregnancy, thereby expressing ‘shared knowledge’.
Example 8. An emoji sequence coded as ‘n/a’.

购物车还是没装满

A loose translation of the text in example 6 is: ‘My shopping cart is still not full’. The emoji sequence merely consists of 6 food-related emoji that function as nouns.

5 Discussion and conclusion

5.1 Research question revisited

The results of the research question – which linguistic elements are formulated in emoji sequences by users to enact their personal identity – support the assumption of this study that emoji sequences provide individuals with a new way of self-presentation in the social media space. These findings differ from how textual and visual self-presentation is discussed in the current CMC literature, to some extent. [17] found that to enact personal identity through text-based profile statuses, WhatsApp users prefer to convey engagement more than stance. This result aligns with [23]: Teens construe their visual selves as invitations to interact with others on social media. This study, however, provides different observations. Its overall finding suggests that Weibo users prefer to present themselves through expressing their own ‘voice’ or community recognized personality (that is, stance) more so than through acknowledging the presence of their audiences and

encouraging them for discourse participation (that is, engagement). This preference of expressing stance is consistent with previous studies which found that the most prevalent speech act conveyed through emoji sequences is making a statement (either subjective or objective) [11, 12]. Furthermore, this study found that, depending on the situational context, the users studied used certain emoji to convey a variety of stance-related and engagement-related linguistic elements (e.g. 'self-mention', 'reader mention'). These findings support the observations that emoji sequences can function as parts of speech and convey a complete message (though it might not be a complex one) [29, 35, 39, 12].

It is unsurprising that the most frequently employed linguistic element identified by this study is attitude markers or the expression of feelings and opinions. This result is consistent with the findings of many previous studies that emotional expression is one of the main aspects of self-presentation on social media [26] and that emoji are semantically rich and emotionally expressive [8], and thus can serve this purpose [26]. Moreover, the use of emoji repetition to convey 'boosters' (that is, emphasizing shared information and marking involvement with the topic) aligns with the statement that repeated emoji are similar to 'beat gestures' in speech [27], and pragmatically function as 'intensifier' to stress and amplify users' utterances [38, 39].

The identification of 'self-mention' (the explicit signal (s) of an author's presence in a message) as well as 'reader mention' (a linguistic marker for acknowledging and bringing message readers to the discourse) is perhaps more surprising. These findings go beyond the common consensus that emoji primarily serve as nonverbal cues in CMC and are used to compensate for the absence of facial expressions and body gestures found in face-to-face conversations [2, 34, 31]. Obviously, the use of emoji that function as 'pronouns' in speech context underlines that users (at least in the Weibo data) are developing their own practices to expand emoji use. The above examples of 'self-mention' (example 1 and 2) depict the users' orientation and worldview of life emanating from their personal encounters and experiences which, altogether, gives a glimpse of their identity [20].

This study found that the most used linguistic element for the expression of engagement is 'directives' (i.e. instructing message readers to either perform an action or accept propositions determined by the writer). This finding aligns with one of the communicative functions of emoji sequences identified by [12], that is, the act of 'manipulate'. The reason might be that these colorful, cute and playful symbols can serve as tone softeners [8], and facilitate message writers' attempts to make a request in a pleasant manner. Surprisingly, only 5% of the sequences did not lend themselves to interpretation as stance and engagement, indicating that the vast majority of the emoji sequences in the Weibo data function pragmatically like textual utterances. Moreover, this study found 7 out of 9 linguistic elements involved in engagement and stance, suggesting that the range of emoji sequences that enact personal identity is limited compared to text. Last but not least, while [9] found that users tend to employ well-crafted textual messages to present an ideal self in CMC (such as avoiding grammatical errors and misspellings), the identified n/a category and the ambiguous sequences that this study encountered during the emoji translation process suggest that emoji sequence formulation is loose and fluid [12].

5.2 Theoretical contributions

First, the study broadens general CMC literature, and specifically emoji-focused literature, by systematically examining and identifying the ways that users employ emoji sequences as textual utterances to construct their personal identity on Weibo. As the literature review shows, previous CMC research has primarily focused on self-presentation through text and/or photographs; the literature on emoji sequences mainly looked at their pragmatic and communicative functions while their social uses remained unexplored. Second, this study contributes to stance and engagement analysis by applying and adapting it to the social media context. Third, it advances the concept of personal identity by adding emoji as a new and digitally-mediated graphical language. Fourth, the results can inform future research aimed at better understanding the emergent user

practice of emoji use in CMC. Finally, the methodological approach employed here is transferable, meaning this paper's research method can be applied to the study of emoji sequence use in other contexts.

5.3 Practical implications

The findings of this study have a number of implications for emoji design and social media platforms. First, the innovative and complex ways that users employ emoji to express engagement and stance suggest that they strive towards making emoji a useful and effective form of language in CMC. At the same time, such user practices also indicate some constraints on using emoji to facilitate digital communication, and therefore inform emoji designers of what is currently lacking in users' emoji repertoire. For instance, emoji repetition that functions as adverbs ('boosters') and animal emoji that function as pronouns ('self- mention', 'reader mention') suggest that emoji design should consider not only the emotional and playful aspects of emoji, but also their linguistic aspects that can function as parts of speech. Second, the observations on non-linear emoji word orders (examples in section 4.2) illuminate the phenomenon whereby users are developing their preferred word orders, rather than following top-down grammatical rules. Designers of emoji and of social media platforms should consider such emergent user practices to be an important factor when creating emoji and improving platform functionalities. 95% of the emoji sequences in the Weibo data were analyzable, suggesting that enacting personal identity through emoji sequences has become an emergent phenomenon (at least in the Weibo data). In light of this, the third recommendation is to provide preformulated emoji sequence options in the form of automatically generated profile statuses. This would be appreciated by users as an alternative to having to search through emoji menus to construct the desired sequences. The last recommendation derives from the observations of this study as well as claims in the literature. According to [19], emoji that render differently across devices and platforms prevent users from using them, due to the concern of causing misunderstanding. Indeed, one of the interpretive difficulties this study encountered is inconsistencies in the ways emoji are rendered. One solution is to provide definitions. In the data analysis this study found that emoji sequences made up of emoji for which definitions were available on Weibo expressed clearer meanings than emoji without definitions.

5.4 Limitations and future research

As the first attempt at systematically examining emoji sequence use in enabling personal identity, this research of course has its limitations and also suggests fruitful directions for future work. First, the data were primarily coded by the author. While engaging in discussions with a linguistic expert and confirming interpretations with the crowdsourced coders and three emoji consultants added reliability, it was not feasible to conduct interrater reliability measurements. Second, this research analyzed sequences of emoji used by celebrities on Sina Weibo, which may limit the generalizability of its findings, e.g., the celebrities may be more creative emoji sequence creators than other users. However, Weibo is the most influential platform in terms of emoji use in China [11] and thus, can be seen as a best practice case.

Despite these limitations, the findings of this research provide compelling evidence of an emerging complexity of emoji sequence use in CMC, a topic which needs to be further explored in future research. Chinese social media, specifically Sina Weibo, provide an ideal research environment due to the high level of social media engagement through emoji found within, but perhaps further testing of the findings derived from this study in other national contexts, or as found in different social media platforms could be the focus. Such research would shed light on the impact of the affordances and use culture of the platform as well as on the national cultures of the various users. Further, by understanding self-presentation related to emoji sequences, this study has helped define and describe the research domain. Future studies can build on its findings, for example by examining the role of emoji sequences in user interactions and impression management, which would further shed light on emoji's social usages.

REFERENCES

[1] M. W. Bauer and G. Gaskell. 2000. Qualitative researching with text, image and sound: A practical handbook for social research. Sage, USA.

[2] J. B. Bavelas and N. Chovil. 2000. Visible acts of meaning: An integrated message model of language in face-to-face dialogue. Journal of Language and social Psychology. 19, 2. 163-194.

[3] D. M. Boyd and N. B. Ellison. 2007. Social network sites: definition, history and scholarship. Journal of Computer- Mediated Communication, 13, 1. 210-230.

[4] H. I. Chen. 2013. Identity practices of multilingual writers in social networking spaces. Language Learning and Technology, 17, 2. 143-170.

[5] T. H. H., Chua and L. Chang. 2016. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. Computers in Human Behavior, 55, 190-197.

[6] H. Cramer, P. de Juan, and J. Tetreault. 2016. Sender-intended functions of emojis in US messaging. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, New York, 504-509.

[7] G. de Seta. 2018. Biaoqing: The circulation of emoticons, emoji, stickers, and custom images on Chinese digital media platforms. First Monday, 23, 9. DOI: https://doi.org/10.5210/fm.v23i9.9391

[8] M. Danesi. 2016. The Semiotics of Emoji: The Rise of Visual Language in the Age of the Internet. Bloomsbury Publishing. UK.

[9] N. Ellison, R. Heino and J. Gibbs. 2006. Managing impressions online: Self-presentation processes in the online dating environment. Journal of computer-mediated communication, 11, 2. 415-441.

[10] J. D. Fearon. 1999. What is identity (as we now use the word). Unpublished manuscript, Stanford University, Stanford, California. USA.

[11] J. Ge and U. Gretzel. 2018. Emoji rhetoric—A social media influencer perspective. Journal of Marketing Management, 15-16. DOI: 1272-1295. https://doi.org/10.1080/0267257X.2018.1483960

[12] J. Ge, and S. C. Herring. 2018. Communicative functions of emoji sequences on Sina Weibo. First Monday, 23, 11. DOI: https://doi.org/10.5210/fm.v23i11.9413

[13] S. Haysom. 2015. Roger Federer's emoji game is absolutely off the charts. Mashable. https://mashable.com/2015/07/06/tennis-players-emoji/#M_MeRRPXYGqB

[14] S. C. Herring. 2004. Computer-mediated discourse analysis: An approach to researching online behavior. In Sasha A Barab, Rob Kling, and Gray, J. H. (Eds.). 2004. Designing for Virtual Communities in the Service of Learning. Cambridge University Press, New York. 338-376.

[15] S. C. Herring and A. R. Dainas. 2017. "Nice picture comment!" Graphicons in Facebook comment threads. In Proceedings of the 50th Hawaii International Conference on System Sciences. Waikoloa. 2185 –2194. URI: http://hdl.handle.net/10125/41419

[16] S. Kapidzic. and S. C. Herring. 2015. Race, gender, and self-presentation in teen profile photographs. New Media & Society. 17, 6. 958-976.

[17] P. A. A. Herzuah. 2018. Enacting Personal Identity Through Language: An Exploration of Some Profile Statuses of WhatsApp Messenger. Journal of Communications, Media & Society (JOCMAS) 5, 1. 99-119.

[18] T. Highfield. 2018. Emoji hashtags//hashtag emoji: Of platforms, visual affect, and discursive flexibility. First Monday. 23, 9.

[19] H. Hillberg, Z. D. Kluver, L. Terveen and B. Hecht. 2018. What I See is What You Don't Get: The Effects of (Not) Seeing Emoji Rendering Differences across Platforms. In Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 124.

[20] K. Hyland. 2005. Stance and engagement: a model of interaction in academic discourse: discourse studies. SAGE, London.

[21] K. Hyland. 2018. Metadiscourse: Exploring interaction in writing. Bloomsbury Publishing, UK.

[22] A. M. Jaffe. 2009. Indeterminacy and regularization. Sociolinguistic Studies, 3, 2. 229-251.

[23] S. Kapidzic. and S. C. Herring. 2015. Race, gender, and self-presentation in teen profile photographs. New Media & Society, 17, 6. 958-976.

[24] A. M. Kaplan. and M. Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53, 1. 59-68.

[25] R. H. Kolbe. and M. S. Burnett. 1991. Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. Journal of consumer research, 18, 2. 243-250.

[26] C. Maíz-Arévalo. 2018. Emotional Self-presentation on WhatsApp: Analysis of the Profile Status. Russian Journal of Linguistics, 22, 1. 144-160.
[27] G. McCulloch and L. Gawne. 2018. Emoji grammar as beat gestures. In Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media, Palo Alto, CA.
[28] J. Miller. 2004. Identity and language use: The politics of speaking ESL in schools. A. Pavlenk and A. Blackledge (Eds) Negotiation of identities in multilingual contexts, Multilingual Matters LTD, UK. 290-315.
[29] J. Monti, F. Sangati, F. Chiusaroli, M. Benjamin, and S. Mansour, 2016. New online tools and digital environments for translation into emoji, In Proceedings of Third Italian Conference on Computational Linguistics. Accademia University Press, Torino, Italy, 211-215.
[30] B. Norton. 2013. Identity and language learning: Extending the conversation. Multilingual Matters, Toronto.
[31] P. K. Novak. J. Smailović. B. Sluban, and I. Mozetič. 2015. Sentiment of emojis, PloS One, 10, 12. 1–22.
[32] J. H. Park, P. Xu and P. Fung. 2018. PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and# hashtags, arXiv preprint arXiv:1804.08280, at https://arxiv.org/pdf/1804.08280.pdf
[33] J. Paliszkiewicz and M. Sawicka. 2016. Impression management in social
media: The example of linked. Management, 11, 3. 203-212.
[34] U. Pavalanathan and J. Eisenstein. 2016. More emoji, less:) The competition for paralinguistic function in microblog writing, First Monday, 21, 11, at http://firstmonday.org/ojs/index.php/fm/article/view/6879/5647
[35] W. Radford, A. Chisholm, B. Hachey, and B. Han, 2016. Telephone. Person. Sailboat. Whale. Okhand or "Call me Ishmael" - How do you translate emoji? Cornell University Library. doi: arXiv preprint arXiv:1611.02027, at https://pdfs.semanticscholar.org/acf0/e1c8d8eacdc6d62d20cc00db712f7f8a6994.pdf
[36] A. Robertson. W. Magdy and S. Goldwater. 2018. Self-Representation on Twitter Using Emoji Skin Color Modifiers. arXiv preprint arXiv:1803.10738, at https://arxiv.org/pdf/1803.10738.pdf
[37] G. Salimkhan. A. M. Manago and P. M. Greenfield. 2010. The construction of the virtual self on MySpace. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 4, 1, at https://cyberpsychology.eu/article/view/4231/3275
[38] A. Sampietro. 2016. Exploring the punctuating effect of emoji in Spanish WhatsApp Chats. Lenguas Modernas, 47. 91-113.
[39] C. Schneebeli. 2017. Emoticons and/as linguistic cues in CMC: the example of YouTube comments to "DIY makeup tricks with school supplies (Miranda Sings). In Proceedings of VINM 2017: Visualizing (in) the new media. Neuchâtel, Switzerland.
[40] L. A. Shafie. S. Nayan. & N. Osman. 2012. Constructing identity through Facebook profiles: online identity and visual impression management of university students in Malaysia. Procedia-Social and Behavioral Sciences, 65, 3. 134-140.
[41] A. Siibak. 2009. Constructing the self through the photo selection-visual impression management on social networking websites. Cyberpsychology: Journal of psychosocial research on cyberspace. 3, 1. https://cyberpsychology.eu/article/view/4218
[42] G. Stald. 2013. Mobile identity: Youth, identity, and mobile communication media In D. Buckingham. (Ed). Youth, identity and digital media. Cambridge M.A: MIT Press. UK.
[43] M. M. Strano. 2008. User descriptions and interpretations of self-presentation through Facebook profile images. Cyberpsychology: Journal of Psychosocial Research on Cyberspace. 2, 2. https://cyberpsychology.eu/article/view/4212
[44] E. G. Stamatov. 2017. Do emoji use a grammar? Emergent structure in non-verbal digital communication. Master Thesis.
[45] K. Steinmetz. 2014. Time Exclusive: Here are rules of using emoji you didn't know you were following, Time, at http://time.com/2993508/emoji-rules-tweets/
[46] S. Sugiyama. 2015. Kawaii meiru and Maroyaka neko: Mobile emoji for relationship maintenance and aesthetic expressions among Japanese teens. First Monday. 20, 10.
[47] K. H. Tan. and S. Dras. 2015. What's on your mind? self-presentation and impression management among Malaysian tertiary level students, at http://www.researchgate.net/publications/290437681

[48] B. I. T. Al-Saleem. 2011. Language and identity in social networking sites. International Journal of Humanities and Social Sciences. 1, 19. 197-202.
[49] K. Tracy and J. S. Robles, 2013. Everyday talk: Building and reflecting identities. Guilford Press. New York.
[50] R. Zhou. J. Hentschel. and N. Kumar. 2017. Goodbye text, hello emoji: Mobile communication on WeChat in China, In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 748-759.

Multimodal Emotion Classification

Anurag Illendula
Department of Mathematics
Indian Institute of Technology Kharagpur
Kharagpur, India
aianurag09@iitkgp.ac.in

Amit Sheth
Kno.e.sis Center
Wright State University
Dayton, Ohio
amit@knoesis.org

ABSTRACT

Most NLP and Computer Vision tasks are limited to scarcity of labelled data. In social media emotion classification and other related tasks, hashtags have been used as indicators to label data. With the rapid increase in emoji usage of social media, emojis are used as an additional feature for major social NLP tasks. However, this is less explored in case of multimedia posts on social media where posts are composed of both image and text. At the same time, w.e have seen a surge in the interest to incorporate domain knowledge to improve machine understanding of text. In this paper, we investigate whether domain knowledge for emoji can improve the accuracy of emotion classification task. We exploit the importance of different modalities from social media post for emotion classification task using state-of-the-art deep learning architectures. Our experiments demonstrate that the three modalities (text, emoji and images) encode different information to express emotion and therefore can complement each other. Our results also demonstrate that emoji sense depends on the textual context, and emoji combined with text encodes better information than considered separately. The highest accuracy of 71.98% is achieved with a training data of 550k posts.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Social media*;

KEYWORDS

Emoji Understanding, Emotion Classification, Multimodal Analysis

ACM Reference Format:
Anurag Illendula and Amit Sheth. 2019. Multimodal Emotion Classification. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308560.3316549

1 INTRODUCTION

Emotion is any conscious experience characterized by intense mental activity and a certain degree of pleasure or displeasure. It primarily reflects all aspects of our daily lives, playing a vital role in our decision-making and relationships. In recent years, there have been a growing interest in the development of technologies to recognize emotional states of individuals. Due to the escalating use of social media, emotion-rich content is being generated at an increasing rate, encouraging research on automatic emoji classification techniques. Social media posts are mainly composed of images and captions. Each of the modalities has very distinct statistical properties and fusing these modalities helps us learn useful representations of the data [3]. Emotion recognition is a process that uses low-level signal cues to predict high-level emotion labels. With the rapid increase in usage of emojis, researchers started using them as labels to train classification models [6]. A survey conducted by secondary school students suggested that the use of emoticons can help reinforce the meaning of the message [1]. Researchers found that emoticons when used in conjunction with a written message, can help to increase the "intensity" of its intended meaning [5].

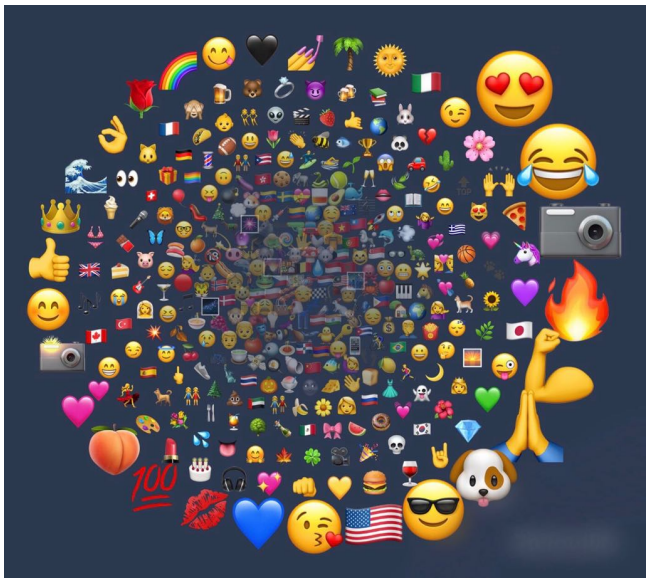


Figure 1: Spiral representing usage of different Emojis on Instagram. The image is copied from https://bit.ly/2W3ks5u, Article by Stefan Pettersson

Emojis are being used for the visual depictions of human emotions [23]. Emotions help us to determine the interactions among human beings. The context of emotions specifically brings out the complex and bizarre social communication. These social communications are identified as the judgment of other personś mood based on his emoji usage [23]. According to a study made by Rajhi et al. [23], the real-time use of emojis can detect the human emotions in different scenes, lighting conditions as well as angles in real time. Studies have shown that emojis when embedded with text to express emotion make the tone and tenor of the message

[1] https://bit.ly/2NyyiIp

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316549

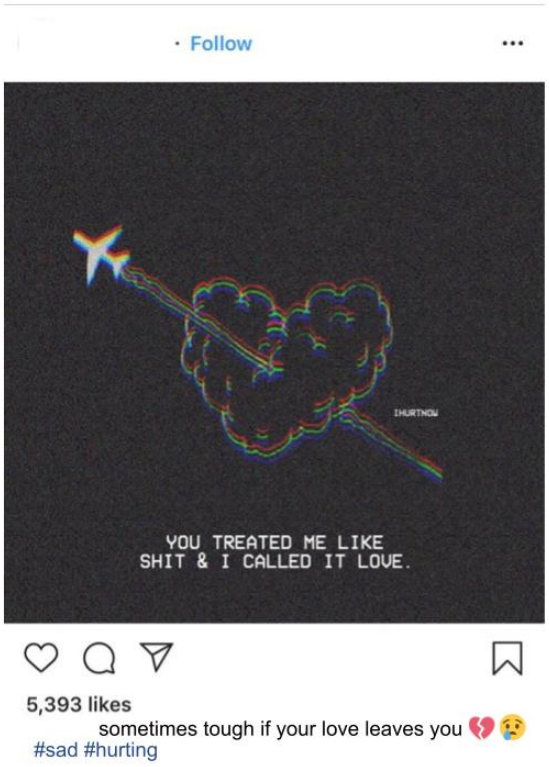


Figure 2: Example of an Instagram post which belongs to "Sad" emotional category in our dataset

clearer. This further helps in reducing or eliminating the chances of misunderstanding, often associated with plain text messages[2]. Recent study proved that co-occurrence helps users to express their sentiment more effectively[3] [13].

Psychological studies conducted in the early '80s provide us strong evidence that human emotion is closely related to the visual content. Images can both express and affect peopleś emotions. Hence it is intriguing and important to understand how emotions are conveyed and how they are implied by the visual content of images. With this as a reference, many computer scientists have been working to relate and learn different visual features from images to classify emotional intent. Convolutional Neural Networks (CNNs) have served as the baselines for major Image processing tasks. These deep CNNs combine the high and low-level features and classify images in an end-to-end multi layer fashion.

Earlier most researchers working in the field of social NLP have used either textual features or visual features, but there are hardly any instances where researchers have combined both these features. Recent works by Barbieri et al.'s [3], Illendula et al.'s [12] on multimodal emoji prediction and Apostolova et al.âĂŹs [1] work on information extraction fusing visual and textual features have shown that combining both modalities helps in improving the accuracies. While a high percentage of social media posts are composed of both images and caption, the researchers have not looked at the multimodal aspect for emotion classification. Consider the post in Firgure 2 where a user is sad and posts an image when a person close to him leaves him. The image represents a disturbed heart and has a textual description "sometimes tough if your love leaves you #sad #hurting" conveys a sad emotion. Similarly the emoji used 💔, 😢 conveys the emotion of being depressed. We hypothesize that all the modalities from a social media post including visual, textual, and emoji features, contribute to predicting the emotion of the user. Consequently, we seek to learn the importance of different modalities towards emotion prediction task. We first discuss relevant research in Section 2. The development of our dataset and present preliminary data analysis experiments which illustrate the importance to study the usage of emojis in different emotional contexts is described in Section 3. We present our model and approach of multimodal emotional classification in Section 4. The results from different experiments and relevant discussion are in Section 5 and Section 6, respectively. We conclude with some interesting findings and our future plan in Section 7.

2 RELATED WORK

Most NLP tasks are limited to the scarcity of labeled data. Earlier many researchers have used manual annotation technique to evaluate their models, but this requires much understanding of the emotional content of all the expressions, is time-consuming, requires much effort and may differ according to one's perspective. This creates a misinterpretation of emotion and effects the accuracies of respective tasks. Hence in most social NLP related tasks namely sentiment analysis and other emotion classification tasks hashtags were used as features for automatically labeling data to corresponding categories [6, 30]. However, the rapid increase in the usage of emojis on social media helped researchers use emoticons as features for data labeling. Felbo et al. [6] has introduced a transfer learning approach for emotion classification, sentiment analysis, sarcasm identification through emoji prediction on the text.

Using emoji knowledge for sentiment analysis, emotion classification and related tasks is not a new idea. Researchers have started using Emojinet [31] to learn embeddings for sentiment analysis task and have achieved better accuracies than the previous state of the art emoji embeddings [32]. There have been many approaches which use emoji as a feature to classify sentiment on social media using emoji as a feature, Illendula et al. [13] has used emoji co-occurrence has a feature to learn sentiment features of emojis. Also, emojis have also been a very important feature to classify emotional content; previous research has always manually specified which emotional category each hashtag or emoji belongs to [22]. Prior research work has used theories of emotion such as Ekman's six basic emotions, Pluchtik's eight basic emotions, and other psychology works [21, 26, 28]. Shaver et al.'s [26] work which is the most used and most cited work helped to automatically label tweet without any human intervention to take care of biases and misinterpretation to corresponding emotional label and develop an emotion classification model for text [30].

Content-Based Information Retrieval (CBIR) is the historical line of research in multimedia tasks. This task usually deals with retrieving images in the dataset that are most similar to the query image. Bag-of-Words representation [34] has seen a sustained line

[2]https://bit.ly/2AV1sNA
[3]https://bit.ly/2QW91sx

Table 1: Emotion words used for collecting tweets and the statistics of instagram posts in our dataset after filtration

Emotion	# of Hash Tags	Examples of Hash Tags	# of Instagram Posts with Emoji	# of Instagram Posts without Emoji	Number of Instagram Posts	Percentage of Posts with Emoji
Anger	23	aggravation, irritation, agilation, anger	59458	28549	88007	67.56
Fear	22	shock, fear, fright, terror, panic	44511	28078	63644	61.32
Joy	36	excited, happy elated, proud	55566	29947	85513	64.98
Love	17	affection, love, loving, fondness	54156	31596	89486	56.38
Sad	36	sorrow, unhappy, depressing, lonely	54156	31596	85752	63.16
Surprise	5	amazement, surprise, astonishment	45640	17230	72870	62.25
Thankful	2	thankfulness, thankful	42870	31308	74178	57.80
Overall	141		352649	205746	558395	63.16

of research in this task which has been effective up to million sized image datasets. This Information Retrieval task has eventually led to research in image classification which is one of the fundamental challenges in Computer Vision. Convolutional Neural Networks (CNNs) [16] have shown promising results in image classification. Research on shortcut connections has been an emerging topic since the development of multilayer perceptrons which has shown promising results for Image processing tasks. Generally, these multiple layers have been connected using shortcut connections using gated functions [10]. In image classification, depth of the network, i.e., number of layers within the network is of crucial importance as noted by Simonyan et al. [27]. Increasing the depth can have an adverse effect on the image classification task. Most notably, vanishing/exploding gradients problem [7] and the degradation problem [8] are of significant importance. These problems are overcome by the introduction of shortcut connections and residual representations introduced in Residual Networks [9] which won the 1st place in ImageNet 2015 Image classification competition [4]. Residual Networks and its extensions which consist of many residual units have shown to achieve state-of-the-art accuracy for image classification tasks on datasets such as ImageNet [24].

Knowledge graphs have been proved to be on the important addition to data for training machine learning and deep learning model architectures for various Natural language and Image Processing tasks. Recent research in NLP Emojinet has improved the accuracies of emoji similarity [32], learning emoji embeddings, emoji sense disambiguation [31]. Emojinet has also improved the accuracies of emoji prediction in case of images [12]. In the case of image processing tasks, knowledge graphs have also improved the task of video captioning [29] and other related tasks [19].

In this paper, we present an emotion classification approach using techniques in deep learning leveraging the three modalities from social media posts namely textual, visual and emoji features. We use state of the art approaches for learning features from three modalities namely ResNet model architecture, attention mechanism, Bag of words model for visual, textual and image features respectively. We used two different word embeddings namely FastText (word embedding model which is capable of capturing sub word information) [14] model trained on the set of the processed captions (we explain the pre-processing steps involved in Section 3) and pretrained fast text model trained on Wiki corpus [20]. We use the bag of words model developed by Wijeratne et al. [32] to learn the emoji embeddings using the emoji knowledge concepts extracted from Emojinet. We report the results observed considering different emoji knowledge concepts from EmojiNet namely emoji names, emoji senses, emoji sense definitions, and three different word embedding models. We also discuss the importance of different modalities towards emotion classification and report our results and observations in further sections.

3 DATASET DEVELOPMENT

3.1 Instagram Dataset

Instagram is one of the major platforms where people share multimedia posts composed of images and text in the form of description. Hence Instagram provides us an emotion-rich content to analyze different user emotions using machine learning and deep learning techniques. We have used the same approach as Wang et al. [30] to collect labeled data. Following Shaver et al. [26] they collected set of possible hashtags for six different emotions (for example words amazement, surprise, astonishment convey the emotion "Surprise"). This approach to collect emotion-labeled posts has been proved to be effective to label text. Shaver et al. have constructed an emotion-related tree where the first layer represents the parent category, i.e., the six different emotion, the second layer consists of 25 emotion-related keywords categorized into six different parent emotions. We have added a seventh emotional category different from the above mentioned emotional categories named "Thankful". We then checked posts having hashtags #gratified, #humbled, #blessed, #thankful on instagram, because in general these key words mean thankful according to english literature. But there are

[4] https://bit.ly/2y4J8Cz

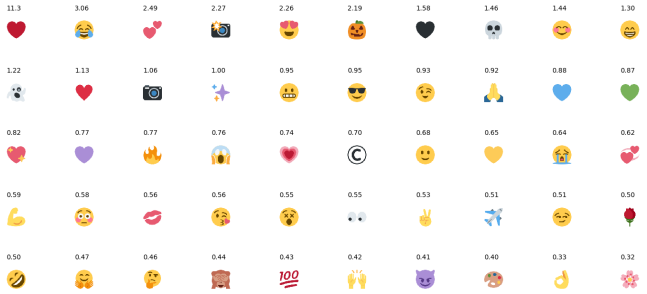


Figure 3: Percentage of number of posts having different types of emojis embedded in their post description.

hardly any posts having a hashtag #gratified (around 4k posts in total and no posts in our analysis period). We then checked posts having hashtags #humbled and #blessed but several posts convey a Joy emotion rather than Thankful emotion. Prior research of study of emotion related key words convey that # blessed could be considered as happy emotion but not necessarily Thankful or Joy or Love emotion is particular [18]. Hence we considered only the key words reported by Wang et al. for the seventh emotional category namely #thankful and #thankfulness for Thankful emotion. We then followed the pre-processing steps for emotion-related keywords by adding lexical variants of a word and removing ambiguous words. Table 1 gives a complete idea of the set of hashtags along with the number of posts collected from Instagram. We then used Instalooter[5] API which is an open source tool to collect public posts from Instagram. Instalooter API gives us access to HashTaglooter method similar to Twitter's hashtag search which helps us to collect multimedia posts having a particular hashtag. For example, we collected the posts having "#amazement" and labeled the post to Surprise emotional category. Similarly, we extracted the emotion-labeled multimedia posts for other emotion labels.

3.2 Pre Processing

In this section, we explain the different pre-processing steps to ensure the quality of our Instagram dataset. In total, we have collected about 1.1 million posts of different languages from Instagram using instalooter API posted between 1st December 2018 to 20th December 2018, a duration of 20 days. The preprocessing of this corpus consisted of a set of filtering, followed by annotation, as discussed next.

Filtering:

- Our current approach is limited to English language only, so we filtered the dataset for the posts that used English language only.
- We also ensured that the multimedia posts contain only images but not videos. Instagram also allows users to post multiple images in the same post, so we have filtered them and also considered posts containing one image but not multiple images.
- The next step of filtration involved removal of posts which have multiple hashtags belonging to different emotional content. For instance, user's have used surprise and joy in the same post, and we couldn't categorize this post to a single emotional category.
- We also made sure that we remove all the emotion-related hashtags, for instance, users may have used #amazement and #surprise in the same post. Hence we removed all the emotion-related hashtags which are present in our lexicon.
- We then removed posts having URL's embedded in the textual description section. We hypothesize that the content obtained from the URLs are likely to be important for understanding the emotional content of the post.

Annotation and Text Processing:

- Each post was annotated first by removing the emotion-related keywords from the textual description with the corresponding emotion label.
- Suppose, if the description of the posts has #bliss, then we removed the emotion-related hashtag and then labeled the post to "Joy" emotional category, if a hashtag is not an emotion-related keyword we converted the hashtag to corresponding word ("#car" to "car").
- We also followed some text pre-processing steps like removing characters repeated more than two times in a word (e.g., looooool to lol).
- We replaced user mentions to "USER" to anonymize users.

3.3 Explorational Data Analysis

Leveraging emoji as an additional feature has improved accuracies of various socialNLP tasks. Examples include improved user classification based on marijuana usage [17], better twitter street gang member identification [2], and higher accuracy in sarcasm and emotion identification on Twitter [6]. A recent study by Scott Ayres on emoji usage on the Instagram platform reports that 31% of all image posts and 29% of all video posts contained emojis [6]. Hence, we have analyzed the usage of emojis in each emotional category. Table 1 represents the number of posts having emojis embedded in the textual description out of the complete set of posts under each emotional category. We observed that about 63.16% of posts contain emojis. In many cases, users use them to express their emotion which cannot be conveyed in the form of text. Hence we explore the use of domain knowledge of emojis from EmojiNet[7] in our multimodal emotion classification task and reported our results in further sections.

Figure 3 shows the percentage of posts having different emojis. We have found that users used more than 1800 different types of emojis in their posts, the emoji ❤️ which is the most frequently used emoji, is used in only 11.3% posts, which is about $1/9^{th}$ of the total number of posts. We have then looked into the use of emojis to express a particular emotion. Table 2 reports the five most frequently used emojis to express a particular emotion. Consequently we looked at the sense forms of these emojis from EmojinNet, and the results looked convincing. For instance, the most frequently used emojis under the emotional category "Sad", 💔, 😔, 😭, 😢, 🥀

[5]https://bit.ly/2RGBYxF

[6]https://bit.ly/2E7rO1L

[7]https://bit.ly/2HiSRIq

Table 2: The five most frequently used emojis to express a particular emotion

Emotion \ Emojis	1	2	3	4	5
Anger	🖤	🔥	✨	😠	💔
Fear	🎃	👻	💀	😱	😬
Joy	❤️	😍	😊	💜	🎵
Love	❤️	😍	💕	😘	💙
Sad	💔	😔	😭	😢	🥀
Surprise	😲	😍	❤️	😊	✨
Thankful	❤️	💕	🙏	😊	💗

have sense forms which are closely related to sad or depressed. Similarly, we have studied emoji usage in other emotional categories and observed that the emojis corresponding to each emotional category have sense forms related to the emotional category. This supports our hypothesis that emojis play a vital role in relating to users' emotion.

4 PRE TRAINING

We collected set of all user captions from each post and followed pre-processing steps and trained a fastText word embedding model. We chose fasttext over other conventional word embedding models due to its capability of capturing sub-word information which plays a significant role in social NLP tasks. Also fasttext word embedding model has proved to give greater accuracies compared to other word embedding models over various NLP tasks such as sentiment analysis, emotion detection, sarcasm identification [14], and emoji prediction [12].

As we have observed that the emotional intent of the emoji used and the emotion expressed by the user in the post are parallel, we tried to use EmojiNet which lists 21,000 emoji sense definitions and 4,618 emoji sense forms to 2,389 emojis. Emoji sense forms list the possible sense forms related to each emoji and emoji sense definitions explaining the context of use of emojis [31]. We chose to use the Bag of words model developed by Wijeratne et al. [32] to learn the emoji representations in the same vector space as words.

5 MODELS

In this section, we present and motivate the models that we used to learn features from an Instagram post composed of a picture and the associated comment for emotion classification task.

5.1 ResNets

ResNet architecture [9], composed of convolution layers is currently the state of the art for image classification tasks achieving high accuracies[24]. Before the introduction of ResNet's, deep CNNs were used for most image processing tasks. ResNet is a feed forward CNN that utilizes two or more convolution layers [25]. It has been observed that the accuracy of image classification majorly depends on the depth of the network, i.e., the number of layers which can be related to the number of layers within the network [7]. Hence

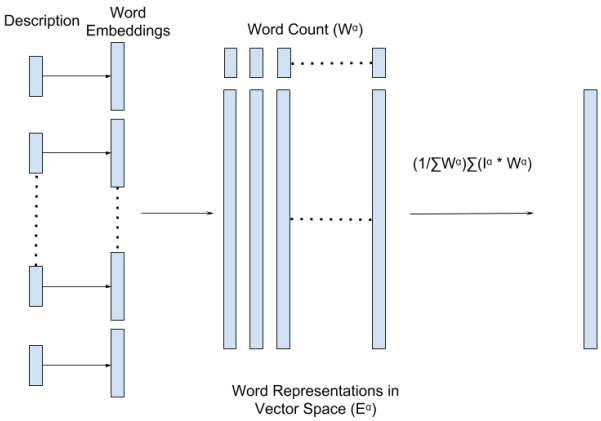


Figure 4: Bag Of Words model to learn emoji representation in the similar vector spaces as words.

we used different types of ResNet model architectures, namely ResNet-101 and ResNet-152 to check the classification accuracy in different cases. ResNet model architecture was first tested on the ImageNet dataset where the number of classes for classification was 1000. Since our task of emotion classification requires us to learn features from both text and image, we use the ResNet model to learn a feature vector from an image. We use the implementation of ResNet by Kotikalapudi which is open-sourced and available on github[8]. Table 3 reports the precision, recall, macro F1 score observed using different ResNet model architectures.

5.2 Bag Of Words Model

As discussed earlier there are different word embedding models which learn rich representations of words in vocabulary using neural networks. The neural network takes a large n-dimensional vector for each word (where n is the number of distinct words in the vocabulary), then learns a transformation of the vector in a low dimensional vector space. There have been many word embedding models to learn word representations including GloVe, skip-gram and CBOW. Similarly, there are different approaches to learn emoji embeddings that use co-occurrence feature [13], skip-gram word embedding architecture [4], and semantic knowledge of emojis from EmojiNet [32]. Since EmojiNet gives us access to the emoji sense forms, we make use of the embedding model developed by Wijeratne et al. [32] to learn emoji embeddings. We also check the accuracy of our approach using other emoji embeddings. Wijeratne et al. [32] replaced the word vectors of all words in the emoji definition and formed a 300-dimensional vector performing vector average. Also, the vector mean (or average) adjusts for word embedding bias that could take place due to certain emoji definitions having considerably more words than others has been noted by Kenter et al. [15].

We make use of different knowledge concepts, namely sense forms, sense definitions, emoji names from EmojiNet and use the bag-of-words model (Figure 4), fastText trained word embeddings to

[8]https://bit.ly/2W69fRC

learn the emoji representations. Since we use the word embeddings to learn emoji representation, we could say that both emojis and words are embedded on a similar vector space. We define two types of knowledge embeddings termed as Emoji_Embeddings _Definitions, Emoji_Embeddings SenseLabels learned using emoji sense definitions and emoji sense forms extracted from EmojiNet respectively. Then we evaluate our model using these embeddings as external knowledge concepts.

Emoji_Embeddings _Definitions: Emoji Definitions are the textual descriptions that relate to the context of use of particular emoji. The emoji embedding from the set of descriptions is calculated using the bag-of-words model shown in Figure 4. For example, consider the emoji 😎, EmojiNet lists "One of the temperate seasons, Summer is the warmest of the four temperate seasons, falling between spring and autumn.", "The period or season of summer." and so forth as emoji definitions for the emoji (😎).

Emoji_Embeddings _SenseLabels: Emoji Sense labels are the list of different senses what emoji mean in different contexts. The emoji embedding from the set of Sense labels is calculated following the bag-of-words model shown in Figure 4. For example "amusing", "swagger" and so forth are the sense labels listed by EmojiNet for the emoji (😎).

Let C_i represent the word count of word W_i, E_i represent the word embedding of word W_i; then the emoji embedding can be calculated as :

$$Emoji_Embedding = \frac{\sum E_i * C_i}{\sum C_i}$$

5.3 Attention Model

Recurrent neural networks (RNNs) are a class of neural networks that take sequential data or time series data as input and compute hidden state vector at each time step, and these networks make use of the entire history of inputs to compute hidden state vector. A LSTM network is a special class of RNNs which has a memory cell and three gating units. These three gating units – input gate, forget gate, and output gate, allow the model to control what information to add, what information to use from previous history, and what information to use to output from current memory cell, respectively. Each gate, implemented as a logistic function σ, takes input vector and outputs a value between 0 and 1.

Bi-directional LSTM's are special kind of LSTM's which combine two LSTM running in forward and backward directions. The hidden state vector $\vec{h}_t$ is computed by the forward LSTM, $\overleftarrow{h}_t$ represent the hidden state vector computed by the backward LSTM. Then we compute the hidden state vector h_t concatenating both the hidden state vector, $h_t = [\vec{h}_t : \overleftarrow{h}_t]$. These networks have been shown to be efficient for a wide range of NLP tasks and have improved the accuracy over the LSTM networks because the current hidden state vector is computed using the past and future information. The importance of a word is highly context-dependent, i.e., the same word can have a different degree of importance in different context. To incorporate this perspective, we add an attention layer [33] on top of this encoded bidirectional LSTM which helps the model decide the importance of each word for the emotion classification task. For instance, words like "lovely" or "extraordinary" are likely to add more weight to the emotion carried in the text. The attention layer helps the model learn the importance of each word using attention scores.

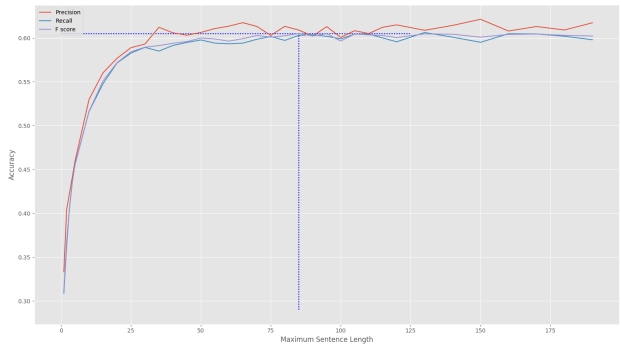


Figure 5: Variation of Accuracy measures vs the input sentence length

Table 3: Accuracies of emotion classification task using visual features from images

Resnet Architecture	Accuracy	Precision	F score
Resnet-101	25.2	25.8	25.4
Resnet-152	**29.6**	**29.8**	**29.7**

6 EXPERIMENTS

In this section, we report the accuracies of different experiments which we performed. We first check the classification accuracy when each modality is used independently, then we combine these modalities and check the classification accuracy of multimodal features towards the emotion classification task. All models are trained using the Keras library and run on Theano background on a cuda GPU, using Adam's gradient descent optimizer. For all experiments, models are trained using 60% of the dataset, validated on 20% of the dataset and tested on the remaining 20% of the dataset.

6.1 Visual Features

We have used the 152-layered residual network which is the current state of the art model architecture for image classification. Also the state of the art model architectures for emoji prediction task for images have used ResNet architecture [3, 12]. Here we trained a 152-layered residual network with learning rate as 0.0001 and stopped the training when there is no further increase in accuracy on the validation set. However, we also checked the classification accuracy using other ResNet model architectures.

6.2 Emoji features

As seen in Table 2, emojis play an integral part in predicting user's emotions on social media. As discussed earlier, we use the Bag of Words model developed by Wijeratne et al. [32] to learn emoji embeddings using emoji knowledge concepts extracted from EmojiNet. We learn four different emoji embeddings using emoji names, emoji

Table 4: Accuracies of emotion classification task using emoji features from caption and fasttext trained word embeddings on the set of descriptions

Knowledge Concept Embeddings	Precision	Recall	Macro F1 Score
Fasttext trained Emoji Embeddings	20.13	19.87	19.43
Emoji Names	27.55	26.67	25.83
Emoji Sense Forms	**34.81**	**30.82**	**32.83**
Emoji Sense Definitions	29.16	27.06	27.38
Processed Emoji Sense Definitions	30.39	28.56	29.66

sense forms, emoji sense definitions, and processed emoji sense definitions. Processed emoji sense definitions are the sense definitions formed after following pre-processing steps, namely removal of stop words, lemmatizing words on the emoji sense definitions extracted from EmojiNet. We consider the vector average of the emoji embeddings of emojis emb2dded in each post as the emoji feature vector for each post. For example if a post is embedded with the emoji set (💔, 💔, 😔), then we take $Feature_Vector$ = 1/3(vec(💔) + vec(💔) + vec(😔)). Next we train a neural network using these feature vectors as input and encoded emotion vector as labels.

6.3 Textual Features

It has been proved that textual information is more appropriate than visual features and increases the accuracy of emoji prediction in case of multimodal information [3]. Hence we tried to check the importance of textual features for emotion classification task. We use the attention mechanism to learn the importance of words towards emotion classification task. Instagram allows users to post a caption of maximum 2200 characters which translates to approximately 300 words [9]. It would be challenging to train a Bi-LSTM with such long input sentence length, we experimented on different input sentence lengths to check the variation of accuracy of Bi-LSTM on sentence length. Figure 5 gives the clear picture of variation of accuracy measures namely precision, recall and F score with input sentence length to the Bi-LSTM using word embeddings trained using fasttext on the set of captions. We have observed that maximum accuracy for classification is achieved when the input sentence length is nearly 80 words per caption and thereafter there is no significant rise in the accuracy of the model. We also reported our accuracies of the emotion classification task using the bag of words model (which is also a extensively used approach for long input sentences) for text using different word embeddings, and we have observed that attention mechanism has achieved better accuracies than the bag of words model. This can be due to the reason that bag of words model gives equal importance to all the words which is not the case with attention mechanism. Table 5 and Table 6 report the accuracies for emotion classification task using different word embeddings and using different model architectures attention mechanism and bag of words model respectively.

[9]https://bit.ly/2RAg2zu

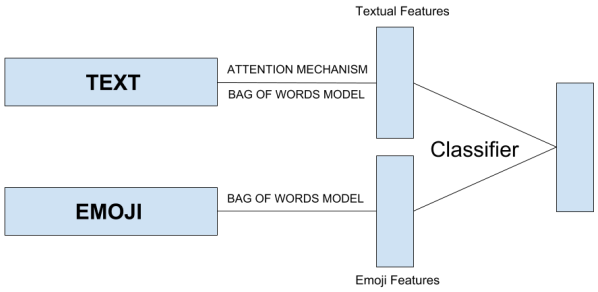


Figure 6: Multimodal emotion classification approach using caption of a post.

6.4 Combining both Textual and Emoji features

Here we present two different model architectures: one which considers complete caption (emojis embedded in between text) as sequential input and fed to a single input layer, while the other considers text and emoji as different input features and are fed to two different input layers as illustrated in Figure 6.

6.4.1 Bag Of Words approach for whole caption. Here we consider complete caption as the input to the classification model, and learn the feature vector by calculating vector average of the embeddings of entities, i.e., word embeddings of words and emoji embeddings of emojis. Since the emoji embeddings are learned using the word embeddings, this supports the addition of word embeddings and emoji embeddings to calculate vector average. Table 8 reports the accuracies of emotion classification task using bag of words model with different emoji knowledge concepts and different word embedding models.

6.4.2 Attention mechanism for whole caption. According to the study by Kevin Cohn which says emojis can be used as language on social media[10] it is noted that the sense of emoji depends on the context of use [31]. Hence Bi-LSTM which have been the SOTA for sequential data would capture the sense of the emoji in the textual context with the help of recurrent units. Hence we train an attention model to check the accuracy where complete caption (emoji+text) is fed to the input layer of a bi-lstm. Table 7 reports the accuracies of emotion classification task using attention mechanism on whole caption.

6.4.3 Considering emoji and text as different features. Recent research by Wijeratne et al. has proved that sense of emoji depends on the textual context [31]. To prove this assertion in the case of emotion classification we use caption without emojis (only text) and emojis embedded in the caption as different features to train classification model. Here we feed the textual input to the attention layer, feature vector learned using emoji embedded in caption to different input layers. We then merge the outputs of attention layer and emoji input layer and add a softmax layer on top of this for classification. The classification accuracies for this model architecture are reported in Table 9. Figure 6 illustrates the model architecture

[10]https://bit.ly/2UbmTRG

Table 5: Accuracies of emotion classification task using textual features from text and attention mechanism at different caption length

Caption Length	Word Embedding Model	Precision	Recall	Macro F1 Score
20	Fasttext trained on Post Descriptions	57.67	57.14	57.17
20	Pretrained Fasttext on Wiki Corpus	41.84	38.92	40.24
40	Fasttext trained on Posts Descriptions	60.59	59.15	59.39
40	Pretrained Fasttext on Wiki Corpus	46.83	42.39	43.20
80	Fasttext trained on Posts Descriptions	61.33	59.72	60.25
80	Pretrained Fasttext on Wiki Corpus	45.95	44.88	43.05

Table 6: Accuracies of emotion classification task using textual features from text and Bag Of Words model

Word Embedding Model	Precision	Recall	F1 Score
Fasttext trained on Posts Descriptions	56.57	55.82	54.50
Pretrained Fasttext on Wiki Corpus	44.26	43.57	43.47

in this case, where we learn independent features and then merge the features.

6.5 Combining textual, visual and emoji features

Barbieri et al. [3] showed that there would be an increase in prediction accuracies if both visual and textual features are combined. We investigated the same in the context of multimodal emotion classification using SOTA model architectures, i.e., ResNet-152 for learning image features, attention mechanism to learn features from text and merge these features to form a hidden layer. We then use a softmax layer on top of this for classification. We report the classification accuracies using this model architecture in Table 10. Figure 7 shows the model architecture.

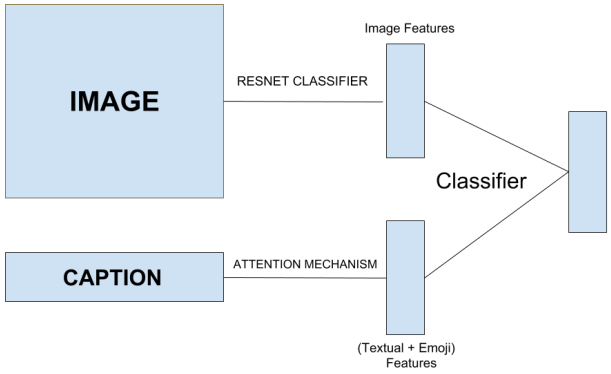


Figure 7: Multimodal emotion classification approach using both caption and image

7 DISCUSSION

We presented extensive experiments to check the accuracies of different approaches and importance of different modalities towards emotion classification. Researchers have identified the importance to study emoji for social NLP tasks considering its usage on social media platforms. Hence we considered emojis as an addition feature to learn the emotional intent from a social media post. We have first looked at how important visual features are to study user emotions. We used ResNet architecture developed by He et al. [8] to learn features from images. We had observed a very low F score of about 30% using ResNet-152 model architecture which is currently the state of the art for image classification. This low accuracy can be explained by the presence of captioned photos which may confuse the model while learning image features for emotion classification [11].

We then looked at the importance of textual features for emotion classification. Recent research has identified the importance of textual features over visual features for task involving social media data [3]. The same has been observed in the case of emotion classification. A caption may contain emoji or text, hence we looked at the importance of these modalities for emotions classification individually. Unlike Twitter where users are restricted to a caption of character length 300, instagram allows users to use a caption of length 2200 which can give room to about 350 words. In our dataset, we have observed that the maximum characters in the caption is nearly 1948, maximum word count of a caption is 423 words, and the average number of words in the caption is about 44 words. We then looked at the variation of accuracies with input sentence length. We observed that the maximum classification accuracy at a input sentence length of 80 words/caption. Further increasing the sentence length did not affect the classification accuracy.

We then looked at the importance of the third modality– emoji features, we used the different emoji knowledge concepts extracted

Table 7: Accuracies of emotion classification task using attention mechanism combining text and emoji as input for Bi-LSTM

Word Embedding Model	Knowledge Concepts	Precision	Recall	Macro F1 Score
Fasttext on Post Descriptions	Emoji Names	66.20	64.82	65.23
	Emoji Senses	65.40	64.96	64.95
	Emoji Sense Definitions	66.04	64.99	65.27
	Processed Emoji Sense Definitions	65.94	64.58	65.04
Pretrained Fasttext on Wiki Corpus	Emoji Names	53.99	52.14	52.42
	Emoji Senses	54.95	51.21	51.99
	Emoji Sense Definitions	52.34	51.85	51.83
	Processed Emoji Sense Definitions	53.72	51.83	52.25

Table 8: Accuracies of emotion classification task using Bag of words technique combining text and emoji as input to learn description embedding using fasttext trained word embeddings on the set of captions

Word Embedding Model	Knowledge Concepts	Precision	Recall	Macro F1 Score
Fasttext on Post Descriptions	Emoji Names	60.51	59.76	59.74
	Emoji Senses	61.29	59.79	60.19
	Emoji Sense Definitions	61.11	60.13	60.50
	Processed Emoji Sense Definitions	60.48	60.03	60.02
Pretrained Fasttext on Wiki Corpus	Emoji Names	54.28	53.85	53.85
	Emoji Senses	53.87	53.48	53.19
	Emoji Sense Definitions	54.20	53.15	53.02
	Processed Emoji Sense Definitions	54.87	53.16	53.51

Table 9: Accuracies of emotion classification task considering text (without emoji) and emoji features as separate input layers

Mechanism for Text	Knowledge Concepts	Precision	Recall	Macro F1 Score
Bag Of Words Approach	Emoji Names	61.38	60.45	60.43
	Emoji Senses	61.94	59.98	60.60
	Emoji Sense Definitions	61.00	60.11	60.19
	Processed Emoji Sense Definitions	61.30	60.63	60.88
Attention Mechanism	Emoji Names	64.68	61.66	62.54
	Emoji Senses	64.71	61.53	62.34
	Emoji Sense Definitions	64.96	61.42	62.12
	Processed Emoji Sense Definitions	63.69	61.54	61.64

Table 10: Accuracies of emotion classification task using attention mechanism combining text and emoji as input for Bi-LSTM

Knowledge Concepts	Precision	Recall	Macro F1 Score
Emoji Names	70.23	69.25	69.42
Emoji Senses	73.79	70.25	71.98
Emoji Sense Definitions	71.56	69.98	70.49
Processed Emoji Sense Definitions	72.23	70.26	70.78

from EmojiNet and trained a neural network. We used the vector average of the emojis embeddings used in each post as the feature vector and input of the neural network. Considering the use of ❤ in different contexts, the emoji ❤ is the most frequently used emoji to express Love or Joy. Hence using only emoji features for emotion classification is not a good idea and the same is observed in the results. The classification accuracy using emoji features is about 32.83% which is very low compared to textual features.

We then looked at the importance of emojis in addition to text for the emotion classification task. We used three different model architectures., The first was the bag of words model for the whole

caption (emojis embedded in between text). The second being attention mechanism with whole caption as input (emojis embedded in between text), where we consider an embedding layer consisting of both textual and emoji embeddings to train an attention model. In the third architecture, emojis and text are considered as two different features and fed through different input layers. The accuracies have been found to be better if emojis embedded with text is considered as sequential input to the attention model. This can be related to a study by Kevin Cohn which says that emojis on social media are being used as a language[11] and hence a Bi-LSTM model would give better accuracies for sequential data. For example consider the caption "My ❤ for life!!", here ❤ is used in context of "love". If both text and emoji are considered as different inputs, this would confuse the model since ❤ can be used in different contexts (as seen in Table 2), this decreases the classification accuracy. This is the reason resulting in high accuracy when both text + emoji is considered as sequential input to attention model.

Finally, we combined the three modalities – textual, visual, and emoji. Table 10 reports the classification accuracies observed using different emoji knowledge concepts wherethe caption (text + emoji) is sent through attention mechanism and image is sent through ResNet-152 model. We then merge the outputs of these layers and train a softmax on top of this for classification. This combination of all the modalities results in better classification accuracy.

8 CONCLUSION AND FUTURE WORK

In this paper, we explored the usage of different emojis in different emotional contexts and the importance of different modalities towards emotion classification task. We have presented a multimodal emotion classification approach which makes use of all modalities – emoji, textual and visual features. We have further shown that combining all modalities can outperform state of the art unimodal approaches (based on only on textual or visual or emoji contents). We also observed better accuracy for the emotion classification task when the caption (emoji and text) is considered as sequential input compared to accuracy when used as different input features. As a future work and with Felbo et al. [6] as reference, we plan to work on building transfer learning approaches using pre-trained classifiers to learn the emotional features from visual and textual contents towards emotion classification task. We also plan to evaluate our approaches using human annotated test set.

REFERENCES

[1] Emilia Apostolova and Noriko Tomuro. 2014. Combining visual and textual features for information extraction from online flyers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1924–1929.

[2] Lakshika Balasuriya, Sanjaya Wijeratne, Derek Doran, and Amit Sheth. 2016. Finding street gang members on twitter. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 685–692.

[3] F Barbieri, M Ballesteros, F Ronzano, and H Saggion. 2018. Multimodal emoji prediction. *arXiv preprint arXiv:1803.02392* (2018).

[4] F Barbieri, F Ronzano, and H Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis.. In *LREC*.

[5] Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2008. Emoticons and online message interpretation. *Social Science Computer Review* 26, 3 (2008), 379–388.

[6] B Felbo, A Mislove, A Søgaard, I Rahwan, and S Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[7] X Glorot and Y Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Yee Whye Teh and Mike Titterington (Eds.), Vol. 9. PMLR, Chia Laguna Resort, Sardinia, Italy. http://proceedings.mlr.press/v9/glorot10a.html

[8] K He, X Zhang, S Ren, and J Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[9] K He, X Zhang, S Ren, and J Sun. 2016. Identity Mappings in Deep Residual Networks. *arXiv preprint arXiv:1603.05027* (2016).

[10] S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[11] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types.. In *Icwsm*.

[12] Anurag Illendula, Kv Manohar, and Manish Reddy Yedulla. 2018. Which Emoji Talks Best for My Picture?. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 514–519.

[13] Anurag Illendula and Manish Reddy Yedulla. 2018. Learning Emoji Embeddings using Emoji Co-occurrence Network Graph. *arXiv preprint arXiv:1806.07785* (2018).

[14] A Joulin, E Grave, P Bojanowski, M Douze, Hé Jégou, and T Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[15] T Kenter, A Borisov, and M de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).

[16] A Krizhevsky, I Sutskever, and G E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017).

[17] U Kursuncu, M Gaur, U Lokala, A Illendula, K Thirunarayan, R Daniulaityte, A Sheth, and I B Arpinar. 2018. " What's ur type?" Contextualized Classification of User Types in Marijuana-related Communications using Compositional Multiview Embedding. *arXiv preprint arXiv:1806.06813* (2018).

[18] Jasy Suet Yan Liew and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*. 73–80.

[19] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844* (2016).

[20] T Mikolov, E Grave, P Bojanowski, C Puhrsch, and A Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *LREC*.

[21] Saif M Mohammad. 2012. # Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 246–255.

[22] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one* 10, 12 (2015), e0144296.

[23] Mohammed Rajhi. 2017. Emotional Recognition Using Facial Expression by Emoji in Real Time. (2017).

[24] O Russakovsky, J Deng, Z Huang, A C. Berg, and L Fei-Fei. 2013. Detecting avocados to zucchinis: what have we done, and where are we going?. In *ICCV*.

[25] Pierre Sermanet and Yann LeCun. 2011. Traffic sign recognition with multi-scale convolutional networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2809–2813.

[26] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology* 52, 6 (1987), 1061.

[27] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[28] Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 121–136.

[29] S Venugopalan, L A Hendricks, R Mooney, and K Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729* (2016).

[30] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 587–592.

[31] S Wijeratne, L Balasuriya, A Sheth, and D Doran. 2017. EmojiNet: An Open Service and API for Emoji Sense Discovery. In *ICWSM*. Montreal, Canada.

[32] S Wijeratne, L Balasuriya, A Sheth, and D Doran. 2017. A semantics-based measure of emoji similarity. In *WI*. https://doi.org/10.1145/3106426.3106490

[33] Z Yang, D Yang, C Dyer, X He, A Smola, and E Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL-HLT*.

[11] https://bit.ly/2UbmTRG

[34] Y Zhang, R Jin, and Z Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* (2010).

Opico: A Study of Emoji-first Communication in a Mobile Social App

Sujay Khandekar
University of Illinois at Urbana-Champaign
khandek2@illinois.edu

Joseph Higgs
University of Illinois at Urbana-Champaign
joebh5@gmail.com

Yuanzhe Bian
University of Illinois at Urbana-Champaign
yb3@illinois.edu

Chae Won Ryu
University of Illinois at Urbana-Champaign
cwryu2@illinois.edu

Jerry O. Talton
Carta, Inc.
jerry.talton@carta.com

Ranjitha Kumar
University of Illinois at Urbana-Champaign
ranjitha@illinois.edu

ABSTRACT

In the last two decades, Emoji have become a mainstay of digital communication, allowing ordinary people to convey ideas, concepts, and emotions with just a few Unicode characters. While emoji are most often used to supplement text in digital communication, they comprise a powerful and expressive vocabulary in their own right. In this paper, we study the affordances of "emoji-first" communication, in which sequences of emoji are used to describe concepts without textual accompaniment.

To investigate the properties of emoji-first communication, we built and released Opico, a social media mobile app that allows users to create *reactions* — sequences of between one and five emoji — and share them with a network of friends. We then leveraged Opico to collect a repository of more than 3700 emoji reactions from more than 1000 registered users, each tied to one of 2441 physical places.

We describe the design and architecture of the Opico app, present a qualitative and quantitative analysis of Opico's reaction dataset, and discuss the implications of Emoji-first communication for future social platforms.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

emoji; social media; mobile app

ACM Reference Format:
Sujay Khandekar, Joseph Higgs, Yuanzhe Bian, Chae Won Ryu, Jerry O. Talton, and Ranjitha Kumar. 2019. Opico: A Study of Emoji-first Communication in a Mobile Social App. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316547

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316547

1 INTRODUCTION

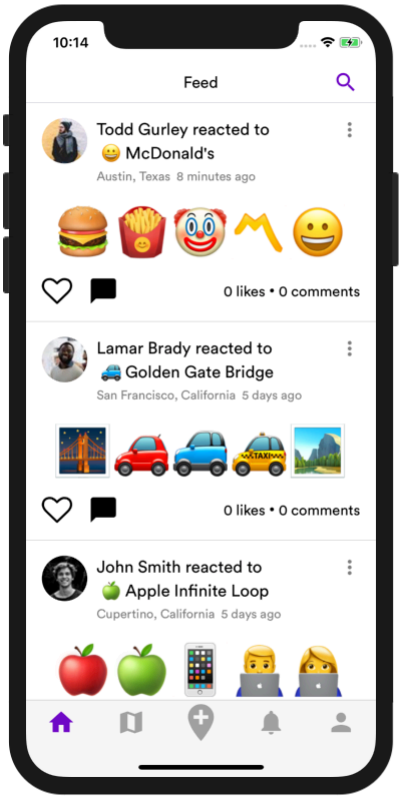


Figure 1: To study the properties of emoji-first communication, we built and deployed Opico, a social media mobile app that allows users to create *reactions* — sequences of one to five emoji — and share them with a network of friends.

Since they were first proposed in 1999 [21], emoji have exploded in popularity, allowing internet users to quickly convey ideas, concepts, and emotions with just a few Unicode characters. These pictorial — often playful — representations offer universal appeal: as of June 2018, the Unicode 11 standard includes 2,823 emoji in its specification [3], and Facebook reports that 92% of users between the ages of 13 and 18 (and 77% of users between the ages of 56 and 64) employ them in communications [2].

Researchers have studied the role emoji play in digital communication, focusing particularly on the way emoji are commonly used to supplement text by emphasizing key concepts or replacing certain words [18]. Some examples exist, however, of people composing messages *entirely* in emoji: Emoji Dick, for instance, is a fanciful sentence-by-sentence translation of Herman Melville's Moby Dick [23].

In this paper, we study the affordances of this kind of "emoji-first" communication, in which sequences of emoji are used to convey meaning without textual accompaniment. Rather than attempt to co-opt existing platforms to conduct such a study at scale in the wild, we built and released Opico, a social media mobile app that allows users to create emoji *reactions* — sequences of between one and five emoji characters — share them with a network of friends, and discuss their meanings (Figure 1).

We released Opico on the Apple App Store and Google Play in early 2018. As of the time of this writing, the app has gained more than 1000 registered users, and generated more than 3700 emoji reactions to a set of more than 2400 physical places. This paper describes the design and architecture of the Opico app, presents a qualitative and quantitative analysis of the resultant dataset, and reports the results of a focused study conducted with 23 Opico users.

Our results suggest that Emoji-first communication represents a powerful and promising mechanism for certain types of digital messages. We demonstrate how the constraints of an emoji-only vocabulary encourage users to succinctly express complex concepts, make jokes, and tell stories. We discuss the implications of the Opico experiment for future social communications platforms, and make a number of data-driven observations about potential ways to evolve the emoji standard.

2 RELATED WORK

A number of researchers have studied emoji communication on various social media and messaging platforms [5, 8, 10, 18, 24, 26, 30]. On these platforms, emoji mainly supplement text, providing affective context, reinforcing concepts found in the text, and/or replacing specific words [5]. Most *emoji sequences* present in this communication comprise repeating emoji, used to emphasize a singular concept [10, 18].

Other studies describe the unexpected affordances of emoji. On Venmo, a peer-to-peer payments platform, people use emoji as shorthand to describe payments. Caraway et al. examined how this emergent emoji use case added a social layer to the app and encouraged users to compose creative sequences of emoji to engage friends in their social network [8]. Zhou et al. studied non-textual communication on WeChat in China, and observed that people use emoji and stickers to converse about sensitive topics and those that are tedious to express in text [30]. Finally, Vidal et al. found that a high percentage of food-related tweets contain emoji, suggesting that they can be an effective representation for describing culinary experiences [26]. This paper examines whether the social, affective, and efficiency affordances of emoji can be amplified in an emoji-first setting, in which users initiate communication with sequences of emoji without textual accompaniment.

While emoji-based communication has many desirable properties, it is more ambiguous than text-based expression [19]. The meaning of emoji is often context-dependent, and tied to the sender's intent [10, 14]. Textual context, however, does not always mitigate miscommunication [20]. Since the artwork for a particular emoji symbol may vary substantially between the major mobile vendors, people's interpretations of emoji may vary across platforms [25]. Emoji usage and interpretation varies across age, gender, geographical location, and social status [6, 9, 16, 29]. This paper is the first to study the interpretability and ambiguity of standalone emoji sequences.

Despite the ambiguity inherent in emoji-based communication, researchers have trained useful models that capture the meaning and sentiment of emoji, suggesting that widely-accepted interpretations of emoji do exist [7, 12, 15, 27, 28]. Many of these models are emoji-word embeddings trained on social media and messaging data, which predict emoji meaning based on textual context [7, 12, 28]. Other models are trained explicitly on crowdsourced annotations for emoji to support retrieval tasks (e.g., emoji keyboard search) [22]. This paper provides evidence that these types of models could be extended to emoji sequences in the future.

3 THE DESIGN OF OPICO

To test the affordances of emoji-first communication, we built Opico, a mobile social app released on the iOS App Store and Google Play. Opico is built to allow users to author and share emoji *reactions* to physical places. Each reaction comprises a sequence of emoji that range between one and five Unicode characters. The five-character maximum provides a Twitter-like constraint, encouraging authors to be succinct and creative, and making reactions easy to parse and consume.

Each reaction is tied to a specific physical location retrieved from the Google Places API [4]. Users can leave a reaction to describe the location itself, a particular experience they've had, or their sentiment about the place. Reactions may be as straightforward as simple one-emoji 👍 (similar to a Facebook "like"), or tell more complicated stories, for instance using 🚘🚦🛬🏃‍♂️😡 to illustrate the experience of being late for a flight due to traffic. Accordingly, Opico serves as a review app, lifelogging platform, and a way to check-in with friends.

3.1 Login and Onboarding

When a user first downloads Opico, they are prompted to authenticate with Google or Facebook account credentials. If the user is new to the platform, the app leads them through a short onboarding sequence that explains Opico's general concept, walks them through the process of creating their first reaction, and helps them find other users to follow. After onboarding, users are taken to the feed, which shows a timeline of recent reactions.

3.2 Posting Reactions

3.2.1 Locations. To create a new reaction, a user must first select a location. Opico displays a map view of nearby businesses and points of interest, and allows users to search by address and by keywords.

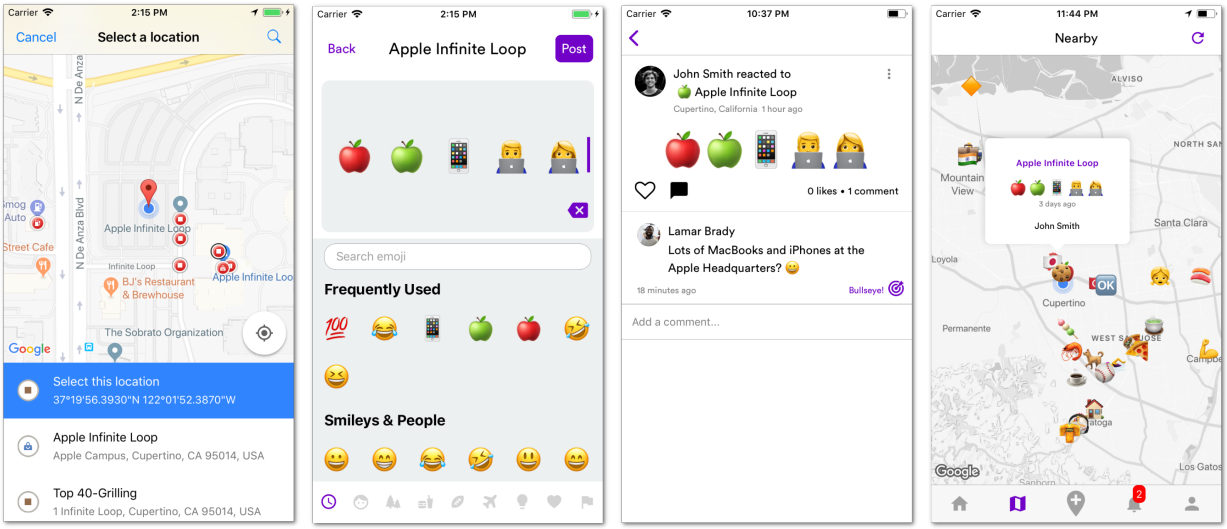


Figure 2: To create a new emoji reaction, a user picks a location and selects up to five emoji. Other users can comment on the reaction in words and earn a "bullseye" 🎯 if the author deems their description correct. In addition to a feed view, Opico provides a map view for viewing reactions.

3.2.2 Keyboard. After a location has been selected, the user is presented with the Opico emoji keyboard. The keyboard presents the entire set of Unicode emoji in a large, scrollable list, divided into a set of curated categories. The bottom row of the keyboard displays quick-jump icons for each category, and the top of the list shows a personalized subset of frequently-used emoji for easy recall.

The keyboard also exposes search functionality so that users can retrieve emoji via keyword. While the Unicode standard provides official names and aliases for each emoji, Opico exposes a broader set of keywords, bootstraped from synonym databases and augmented by logging unsuccessful queries. When a user runs an unsuccessful emoji search, the app logs the query string so that Opico's developers can consider which emoji should have been returned.

3.3 Social Affordances

3.3.1 The Feed. The feed is Opico's primary social view. It shows a list of reactions — sorted in reverse chronological order — from the set of users the logged-in user follows. To ensure a sufficient quantity of fresh content and encourage engagement and discovery, the feed also includes reactions that are "one-hop" away in the network, or from friends of friends. Users can search for new people to follow in the feed's accompanying tab.

3.3.2 Profile & Location. Opico exposes a profile page for each user — which displays a chronological list of their reactions — and a similar profile for each location in Opico's dataset. These views provide a way for a user to track their own travels and recent check-ins, as well as to see a summary of the reactions for a particular place. In addition, user profiles provide links to lists of followers and followees.

3.3.3 Likes, Comments, and Bullseyes. Opico provides a simple ❤️ "Like" mechanism for users to engage with one another's posts, as well as freeform comment functionality for more substantive discussions. Comments in Opico are text fields that are *not* restricted to emoji, and can be used to reply to the content of a reaction, ask questions, or — in the case of ambiguous or difficult-to-interpret posts — guess a reaction's meaning. To encourage users to engage in this latter way, Opico also provides a 🎯 "bullseye" indicator, which the author of a reaction may assign to a single "on-target" comment.

3.3.4 Emoji Maps. Since Opico reactions are location-based, the app also provides a map view of nearby reactions using a Google maps backend. Each location is represented by a single emoji in this view, chosen by picking the most frequently-used emoji character across all user reactions at the location. Tapping on the emoji reveals the most recent reaction for the location, and tapping on the callout takes the user to the reaction itself. The map makes it easy for users to discover locations that have been visited by their network; an alternate, personalized view shows only the user's own reactions.

3.3.5 External Sharing. Opico also allows users to crosspost reactions to their Facebook profile by toggling a button in the posting flow. Crossposts generate a URL with an OpenGraph image displaying the emoji reaction on a map with a small description underneath. This sharing flow allows users to advertise their reactions to friends who have not yet joined the Opico network.

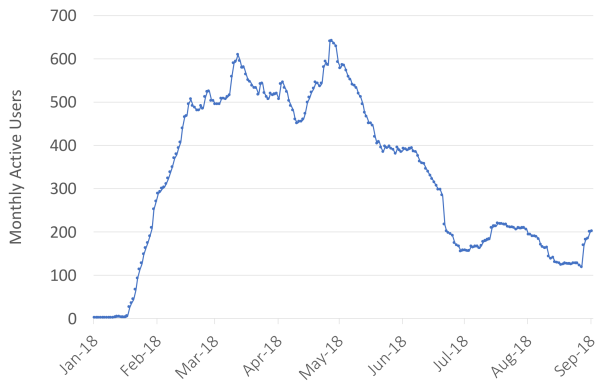


Figure 3: Monthly active Opico users from 1/1/18 to 9/1/18.

4 USERS AND ENGAGEMENT

Opico was released in January 2018 on the Apple App Store and Google Play, and marketed on the campus of the University of Illinois at Urbana-Champaign through posters and word-of-mouth. Since its release, Opico has gained more than 1000 registered users, and generated more than 3700 emoji reactions to more than 2400 physical places.

The platform's registered users are 71.5% male and 28.5% female; 32.98% use Android devices, 67.02% use iOS devices; and 90.73% are from the United States. Most of the users (51.13%) are between the ages of 18 and 24, 33.19% of them are between 25 and 34, 11.39% are between 35 and 44, and 4.3% are between 45 and 54.

Figure 3 shows monthly active users from launch to September 2018, with peak engagement — more than 500 users — between the months of March and May, and usage tapering off during summer break. As of September 2018, users have produced 3769 reactions, given out 6888 likes, left 802 comments, and awarded 126 bullseyes.

On average, each user follows 14.28 ($\sigma = 22.88$) users and is followed by 11.89 ($\sigma = 20.94$) users; 49.74% of the 5826 following relationships are reciprocal.

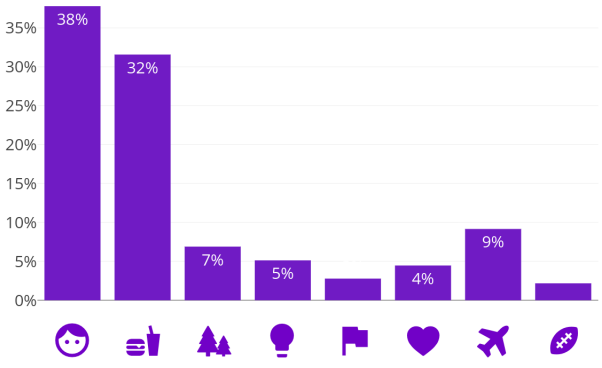


Figure 4: The breakdown of emoji usage by emoji keyboard categories.

Figure 5: Opico emoji reactions often describe activities such as working, traveling, or waiting.

5 USAGE AND REACTION PATTERNS

The 3769 reactions comprise 4.82 emoji on average ($\sigma = 2.17$), which are distributed in the following way across emoji keyboard categories: 38% of the emoji come from the "Smileys & People" category, 32% from "Food & Drink," 9% from "Travel & Places," 7% from "Animals & Nature," 5% from the "Objects," 4% from "Symbols," 3% from "Flags," and 2% from "Activities" (Figure 4).

A significant fraction of emoji used in reactions — "Smileys & People" — describe how people felt at or about a location. "Food and Drink" is the next most popular category of emoji, since most reactions were left for restaurants (58%). The distribution over emoji categories closely matches the distribution over place types: overall, 72% of the reactions were left for food-related places which include restaurants, bars, cafés, and bakeries; and 5% of the reactions were left for airports and lodging. Emoji from the "Animals & Nature" category are often used to describe food, but are also used to describe abstract concepts. For example, the snail 🐌 and turtle 🐢 emoji are commonly used to describe "slow" situations, and the mouse 🐭 is used to describe things that are "small."

Many reactions read like short reviews: since most reactions are associated with food-related establishments, they describe what users ate or drank, and how pleasing the experience was. Other reactions describe activities such as working, traveling, and waiting. Work-related reactions often involve emoji with laptops — 💻, 👩‍💻, 👨‍💻 — and coffee ☕. Travel-related emoji often describe where people are leaving from 🛫, where they are landing 🛬, and "going home soon" SOON 🏠. Finally, waiting-related reactions are often characterized by sequences of people — 🧍🧍🧍 — to indicate lines, and sequences of hourglasses — ⏳⌛ — and clocks — 🕒🕛🕘 — to indicate the passage of time. Figure 5 illustrates many of the common patterns associated with these activities.

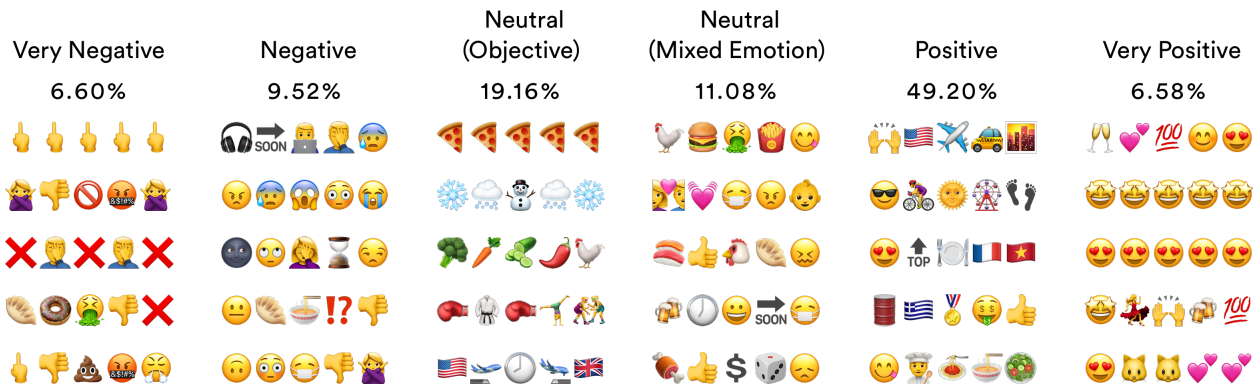


Figure 6: A selection of Opico reactions, categorized by sentiment based on the Emoji Sentiment Ranking [15]. Reactions can be neutral in two ways: they comprise only non-emotive emoji, or they comprise mixed emotions that cancel each other out.

5.1 Emoji Collocation Analysis

To identify patterns and activities in Opico's reaction data, we extracted frequently-occurring emoji bigrams and trigrams, and ranked them by their point-wise mutual information (PMI) [17]. Unlike in the English language, we found very little ordering consistency within consecutive sequences of emoji, so we formed bigrams and trigrams from unordered pairs and triplets, respectively. Since PMI is biased toward infrequent events, we consider only bigrams present in at least five reactions and trigrams present in at least three; additionally, both types of n-grams must have been used by at least two different Opico users to be counted.

Our collocation analysis yielded 248 emoji bigrams with PMIs greater than zero. Many of the bigrams and trigrams are pairs of the same emoji, which is a common pattern in normal emoji usage [18]. Emoji repetition such as 🌽🌽 ($n = 12$) may be used to communicate the plural form of the emoji, or an explicit count. Other times repetition is used for emphasis: 🌶️🌶️ ($n = 7$) and 😀😀 ($n = 8$), may indicate levels of spiciness and happiness, respectively.

Other bigrams pair similar types of emoji, such as those for beer 🍺🍻 ($n = 28$), working on a laptop 👨‍💻👩‍💻 ($n = 19$), and seafood 🦀🦐 ($n = 7$). Or, they bring together commonly-associated items and concepts, such as beer and pool 🍺🎱 ($n = 5$), burgers and fries 🍔🍟 ($n = 57$), coffee and work 💻☕ ($n = 20$), or taking off and landing 🛫🛬 ($n = 8$). Only a few bigrams pair an emotive emoji with a non-emotive one. Most of these bigrams express positive sentiments toward a food item such as pancakes 🍦😋 ($n = 8$), except for one: ⏳😡 ($n = 5$) signifying anger over a long wait.

Perhaps, some of the most interesting bigrams in the dataset are ones that qualify food. When food emoji are paired together, one emoji often describes the main dish, while the other describes the flavor or an ingredient found in that dish. For example, reactions such as 🍫🍰 ($n = 9$) and 🍦🍫 ($n = 7$) signify chocolate cake and ice cream, and ☕🍫 ($n = 11$) represents a 'mocha' or 'hot chocolate.' Similarly, combinations such as 🍔🐔 ($n = 8$) and 🍞🥑 ($n = 8$) can capture dishes such as chicken burger and avocado toast.

Many bigrams pair international flags with food items to describe food from certain countries or cultures: for instance, 🇲🇽🌮 ($n = 10$) and 🇮🇹🍝 ($n = 8$). Many of these bigrams are unsurprising since tacos and pasta are well-known Mexican and Italian dishes. However, some combinations are more unusual, such as 🇮🇳🥞 ($n = 8$). From comments that received bullseyes, we know that this bigram is used in the reaction dataset to communicate the savory Indian pancake-like dish dosa. Since there is no dosa emoji, users repurposed existing emoji to express this new idea.

Overall, we identified 118 trigrams with PMIs greater than zero. Trigrams exhibit patterns similar to bigrams. Many trigrams feature the same emoji repeated three times either to express plurality or emphasis. Reactions such as 😍😍😍 ($n = 21$) or 😠😠😠 ($n = 3$) describe an extra emphasis on an emotion, whereas reactions such as 🧍🧍🧍 ($n = 3$), or 👯👯👯 ($n = 4$) describe "waiting in line" or crowds.

Similar to bigrams, trigrams also bring together similar types of emoji: alcoholic beverages 🍷🍹🥃 ($n = 3$), breakfast foods 🍞🍳🥓 ($n = 3$), and fitness activities 🏃‍♀️🏋️‍♀️💪 ($n = 3$). Trigrams also capture common food dishes, such as a chicken avocado salad 🐔🥑🥗 ($n = 4$), and spicy Indian curry 🇮🇳🌶️🍛 ($n = 3$).

6 SENTIMENT ANALYSIS

Since many emoji explicitly describe a feeling, a reaction's sentiment can be analyzed in a direct manner. We leverage the Emoji Sentiment Ranking [15] lexicon to compute a sentiment score for each reaction in our dataset. The ranking lexicon assigns a sentiment score to each emoji, where positive and negative values correspond to positive and negative sentiment, respectively. Accordingly, emoji with neutral sentiment are assigned a score of 0.

To compute an overall sentiment score for a reaction, we sum up the sentiment scores of its constituent emoji. Based on the sentiment score, we characterize each reaction as having either a strong

positive, positive, neutral, negative, or strong negative sentiment. Reactions can be neutral in two ways: either they comprise only non-emotive emoji, or they comprise both positive and negative emoji that cancel each other out. Figure 6 shows representative reactions from each of the six categories.

The majority of Opico's reactions contain positive sentiments (49.02%). This result is unsurprising given the fact that two most frequently used emoji in the reaction dataset are 😍 or 👍, and suggests that users are often incentivized to react as a way to log and remember their favorite places.

Negative emoji reactions were less common (9.52%), and often described the reason someone was upset: the restaurant was too expensive, or a specific dish was bad, etc. Strong positive (6.58%) and negative (6.60%) reactions comprise repetitive positive and negative emoji, respectively. In these extreme instances, it appears users care more about emphasizing the intensity of their emotion than relating a precise story about why they felt that way.

Users create objective neutral reactions (19.16%) to describe a location without expressing an opinion; such users may use the app to track what they did more than how they felt. 11.05% of the reactions in the dataset capture both positive and negative emotions. For example, the reaction 🐓🍔🤮🍟😋 describes how a user did not like the chicken sandwich at a restaurant, but thought the fries were good.

Our approach for measuring sentiment is not foolproof. For example, some reactions expressed feelings of happiness by using emoji to describe certain idiomatic phrases in the English language such as "so happy I could die" or "feeling like I'm in heaven," with reactions such as 🍣🍲😍💀😇 and 😍🍩➡️🍗😇. These are strong positive reactions, but our scoring approach categorizes both as neutral. Another failure case is irony. For example, the "weary face" emoji 😩 used in the context of describing food can mean "to die for." In the future, we hope that the Opico dataset can be used to build more nuanced, context-sensitive models of emoji sentiment.

7 EFFICIENCY AND AMBIGUITY

One of the primary motivations behind building Opico was to understand whether people could use an emoji-only communication medium to succinctly express complex ideas and stories. A reaction's text translation can help us measure both *efficiency* — how much information can a reaction encode — and *ambiguity* — whether multiple people interpret the message correctly and in the same way. We collected translations through the Opico's in-app comments and bullseye mechanism, and conducted a study with Opico users.

7.1 Bullseyes

Opico's "bullseye" feature allows us to organically collect text-based translations of emoji-only reactions, since bullseyes 🎯 are meant to be awarded to comments that correctly guessed the meaning of a reaction. As of September 2018, users awarded bullseyes to 126 comments. The average length of comments with bullseyes is 26.41 characters ($\sigma = 24.68$), or 4.98 words ($\sigma = 4.67$).

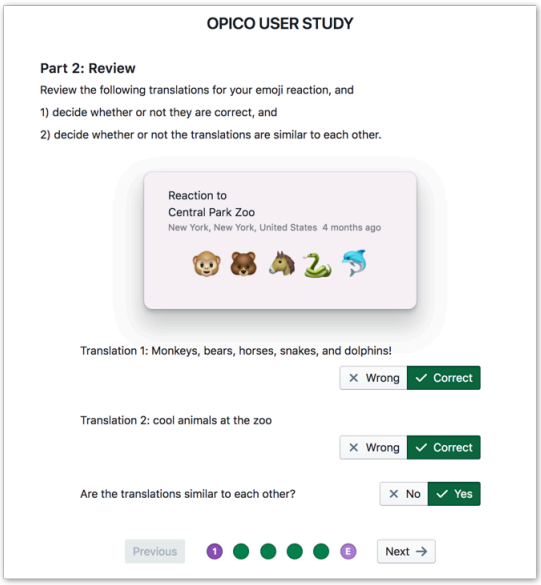


Figure 7: After translating a set of reactions, study participants were asked to review their own reactions and verify whether the translations produced for them were correct.

Creative reactions often elicit comments where other users try to guess the meaning. For example, users were able to guess that the reactions 🐍🎤👸🐍😁 and 🎤🐍💋✨, which were left at stadiums, described Taylor Swift concerts. These reactions allude to the snake emoji's special meaning to Taylor Swift and her fan base. Similarly, users were able to guess that ⚫🐆👍👍👍 and 🐆💎🎮🍿😊, which were left at movie theaters, correspond to the movies "Blank Panther" and "Jumanji." These examples demonstrate how users leverage contextual clues based on location and timeframe to decipher the meaning of reactions.

Many comments that earned bullseyes did not describe the entire reaction, but just the part the author wanted verified. For example, the reaction 🍻📺🏀👏 elicited the comment 'Warriors!', which earned a bullseye despite failing to mention anything about "beer" or "watching the game on TV." Accordingly, we deem the in-app bullseye feature a useful — but incomplete — translation mechanism.

7.2 Study Methodology

To explicitly measure the efficiency and ambiguity of location-based emoji reactions, we recruited participants for a two-part study by emailing the 129 Opico users who had left five or more reactions. The email solicited participation in a user study to better understand how people interpret emoji-based reactions in exchange for emoji swag.

The first part of the study asked users to translate ten emoji reactions written by others into text. We anonymized each reaction, and provided its location and the time it was posted. If a reaction

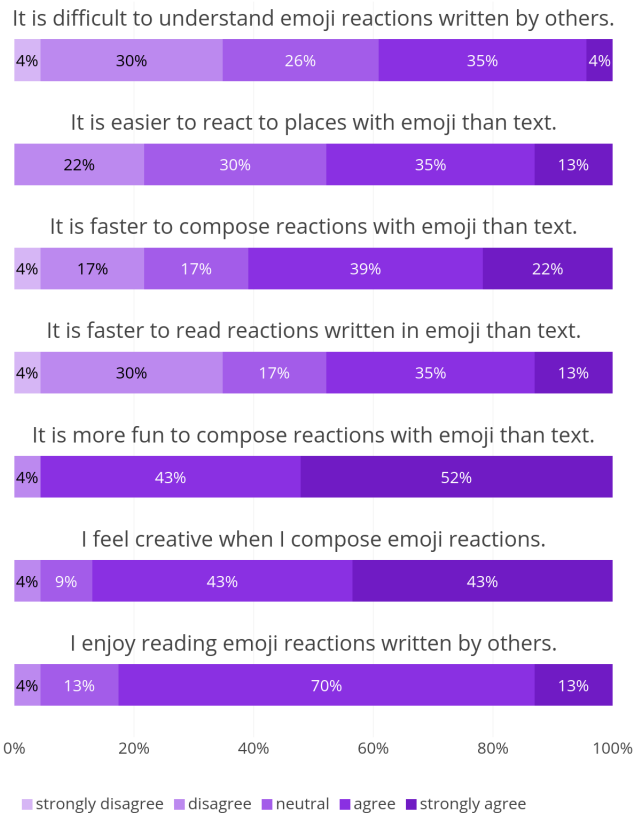


Figure 8: 23 participants were asked to assess statements about emoji reactions on a five-point scale.

was inscrutable, users could select an "I don't know" response instead of composing a translation. To create a representative set of reactions to translate, we randomly sampled five reactions from each participant, and collected two independent translations for each reaction.

After participants completed the first part of the study, we asked them to review their own reactions, and verify whether or not the translations produced by the other participants were correct (Figure 7). In addition, for each reaction with two translations, we asked users to evaluate whether or not the translations were similar to one another. After verifying the translations for their reactions, users completed a survey which asked them to reflect on their experiences composing and consuming location-based emoji reactions. The survey comprised seven questions, and elicited responses on a 5-point Likert-scale, ranging from "strongly disagree" to "strongly agree."

7.3 Study Results

Through our email campaign, we recruited 23 participants. By sampling five reactions from each participants, we created a pool of 115 reactions for participants to translate.

The study's first part produced 219 text translations, and 11 "I don't know" responses. Participants generated two translations for 105 reactions, one translation for nine reactions, and zero translations for one reaction.

In the second part of the study, participants determined that 144 text translations (65.75%) correctly described their own reactions. Out of the 105 reactions with two translations, both translations were correct for 49 reactions (46.67%), and 93 (88.57%) had at least one correct translation.

Additionally, 62 reactions (59.05%) had two similar translations: 45 (72.58%) where both were correct, 6 (9.68%) where both were wrong, and 11 (17.74%) where one translation was right and the other wrong. Some participants indicated that it was possible for reactions to be similar in content and not correctness: for the reaction "🕐🕐🕐😴😩" left at the O'Hare International Airport, the translation "Lots of time waiting at Ohare - sad and sleepy" was marked correct, while the translation "slept in and almost missed flight" was marked wrong.

The average length of the reactions used in the study was 4.10 emoji ($\sigma = 0.99$) and the average length of the resultant text translations was 6.27 words ($\sigma = 3.98$) or 35.15 characters ($\sigma = 21.24$). These results raise interesting questions about whether emoji are more "efficient" than text for conveying information, and in what circumstances.

Figure 8 presents the results of the user survey, which suggests that not all users felt comfortable reading and writing with emoji: 39% of the participants agreed or strongly agreed that "it is difficult to understand emoji reactions," while 22% of the participants disagreed that "it is easier to react to places with emoji than text." Similarly, 21% of the participants disagreed or strongly disagreed that "it is faster to compose reactions with emoji than text," and 34% that "it is faster to read reactions written in emoji than text." Although some users found it difficult to compose and understand emoji reactions, the majority of participants reported that they felt creative (86%) and had fun (95%) while writing them, and that they enjoyed reading them (83%).

8 DISCUSSION AND FUTURE WORK

The results of our study indicate that most location-based emoji reactions are interpretable (since 89% of the reactions received at least one correct translation) and that these interpretations are fairly consistent (59% of reactions were interpreted the same way).

One interesting avenue for future work is studying how ties within one's social network affect one's ability to understand reactions. Users leverage contextual clues based on location and timeframe to decipher the meaning of reactions; to what extent do they also use external knowledge based on personal relationships? Since emoji are often used in personalized ways [29], we wonder how effective emoji-first communication is as a vehicle for broadcasting messages whose true meaning is only understood by those "in the know."

The study also reveals a wide range of comfort with the emoji medium: around half of participants indicated that it is faster to read and write in emoji than in text! As emoji usage increases, we hypothesize that this percentage will also grow.

Below, we sketch a few promising directions for improving emoji-based communication based on Opico: developing better emoji input modalities; informing the emoji specification; and building new, useful emoji interactions.

8.1 Improving Emoji Input Modalities

While Opico's emoji keyboard provides category-based browsing and search, it can still be difficult for someone unfamiliar with the emoji lexicon to select the right emoji. Data collected through informal user interviews suggests that some Opico users did not finish posting reactions because they experienced choice paralysis or were afraid of appearing foolish if they used the "wrong" emoji.

New emoji keyboards can better scaffold emoji novices by suggesting emoji *autocompletions*, or offering templates for longer sequences of emoji based on context [22]. Many messaging apps already suggest an emoji replacement based on the last typed word: this idea could be extended to include longer sequences. On Opico, the keyboard could suggest entire reaction templates based on the type of location, learned from data.

8.2 Informing the Emoji Specification

Although the Unicode lexicon now includes more than 2800 emoji, it still comprises far fewer concepts than the English language. Our informal interviews revealed that some users felt frustrated when they could not find an exact emoji for the concept they wished to convey, while others found this constraint made them more creative. Observing how people creatively repurpose existing emoji to represent new concepts can inform future emoji design.

While most emoji are represented with a single Unicode code point, others are combinations of two or more code points. For example, the 👩‍❤️‍👩 emoji is a combination of the 👩, ❤️, and 👩 code points. The character that joins these code points together is called a Zero Width Joiner, or ZWJ, and, using ZWJ characters with emoji creates Emoji ZWJ Sequences [1]. One benefit to ZWJ sequences is that vendors can introduce their own artwork for any combinations of emoji without first requiring the emoji to be formally ratified by the Unicode consortium.

Opico's dataset — especially the results of our collocation analysis — can help inform future ZWJ sequences that vendors should supply. For example, two frequently used emoji, such as 🥑 and 🍞, can be combined into a single artwork for "avocado toast." The following reactions 🍵🥛⚫, 🌹🍵➕⚫, 🌺🍵⚫😍, ☕🍵⚫⬛, all describe "boba" or "bubble tea," which is a popular drink in the dataset. These reactions suggest that vendors could implement a ZWJ for bubble tea by combining the code points 🍵, 🥛, and ⚫. In the future, apps could also automatically detect semantically meaningful sequences, and leverage visual blending techniques to dynamically generate emoji that express new concepts [11].

8.3 Realizing the Social Affordances

Although the current implementation of Opico allows us to study how users compose and read sequences of emoji, it does not take full advantage of its own representation format. With emoji representations, traditionally complex NLP operations such as summarization and sentiment analysis are greatly simplified. Opico can leverage the emoji representation to easily produce aggregations and visualizations that can reveal interesting patterns about users and places.

For users who utilize the app for life-logging, Opico could provide personalized visualizations that reveal how they spend their time, what kind of food they eat, and so on. Emoji-based search could allow users to quickly run complex queries: a user could find "coffee places to work at" by querying with 💻☕. For users who are looking for a place to eat, Opico could augment its current map visualization to show how nearby restaurants are trending based on sentiment analysis.

Opico reactions could also be aggregated and used to power ranking and recommender systems. The data generated by Opico suggests that an emoji-based rating platform could help overcome the J-shaped distribution that often plagues five-star systems [13]. In contrast to reviews on highly-trafficked e-commerce sites (which are invariably bimodal), sentiments expressed in Opico — where users are more inclined to describe mediocre experiences — follow a more normal distribution. In addition to encouraging users to rate more frequently, emoji-based ratings could provide more descriptive and nuanced feedback for goods, services, and experiences.

REFERENCES

[1] 2016. Emoji ZWJ Sequences: Three Letters, Many Possibilities. https://blog.emojipedia.org/emoji-zwj-sequences-three-letters-many-possibilities/.

[2] 2017. 77% of 56-64 Year Olds Use Emojis On Messenger. https://blog.emojipedia.org/77-of-56-64-year-olds-use-emojis-on-messenger/.

[3] 2018. 157 New Emojis in the 2018 Emoji List. https://blog.emojipedia.org/157-new-emojis-in-the-final-2018-emoji-list/.

[4] 2019. Google Places API. https://developers.google.com/places/web-service/intro

[5] Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling Emoji Popularity Through Semantic Embeddings. In *Proceedings of the 11th International Conference on Web and Social Media*. AAAI Press.

[6] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How Cosmopolitan Are Emojis?: Exploring Emojis Usage and Meaning over Different Languages with Distributional Semantics. In *Proceedings of the ACM Conference on Multimedia*. 531–535.

[7] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. In *Proceedings of the Language Resources and Evaluation Conference*.

[8] Monica Caraway, Daniel A. Epstein, and Sean A. Munson. 2017. Friends Don't Need Receipts: The Curious Case of Social Awareness Streams in the Mobile Payment App Venmo. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing* (2017).

[9] Zhenpeng Chen, Xuan Lu, Wei Ai, Huoran Li, Qiaozhu Mei, and Xuanzhe Liu. 2018. Through a Gender Lens: Learning Usage Patterns of Emojis from Large-Scale Android Users. In *Proceedings of the 2018 World Wide Web Conference*.

[10] Henriette Cramer, Paloma de Juan, and Joel Tetreault. 2016. Sender-intended Functions of Emojis in US Messaging. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM.

[11] João Miguel Cunha, Pedro Martins, and Penousal Machado. 2018. How Shell and Horn make a Unicorn: Experimenting with Visual Blending in Emoji. In *Proceedings of the 9th International Conference on Computational Creativity*.

[12] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[13] Nan Hu, Paul A. Pavlou, and Jie Zhang. 2009. Overcoming the J-Shaped Distribution of Product Reviews. *Commun. ACM* 52, 10 (2009).

[14] Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. 2017. Spice Up Your Chat: The Intentions and Sentiment Effects of Using Emojis. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

[15] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of Emojis. *PLOS ONE* 10, 12 (12 2015), 1–22.

[16] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the Ubiquitous Language: An Empirical Analysis of Emoji Usage of Smartphone Users. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[17] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

[18] Gretchen McCulloch and Lauren Gawne. 2018. Emoji Grammar as Beat Gestures. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media*.

[19] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully happy" or "Ready to fight": Varying Interpretations of Emoji. In *Proceedings of the 10th International Conference on Web and Social Media*.

[20] Hannah Jean Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren G. Terveen, and Brent J. Hecht. 2017. Understanding Emoji Ambiguity in Context: The Role of Text in Emoji-Related Miscommunication. In *Proceedings of the 11th International Conference on Web and Social Media*.

[21] Arielle Pardes. 2019. The Complete History of Emoji. https://www.wired.com/story/guide-emoji/

[22] Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond Just Text: Semantic Emoji Similarity Modeling to Support Expressive Communication. *ACM Transactions on Computer-Human Interaction* 24, 1, Article 6 (2017).

[23] Will Radford, Andrew Chisholm, Ben Hachey, and Bo Han. 2016. 📞👨⛵🛁👌; or "Call me Ishmael" - How do you Translate Emoji?. In *Proceedings of the Australasian Language Technology Association Workshop*.

[24] Channary Tauch and Eiman Kanjo. 2016. The Roles of Emojis in Mobile Phone Notifications. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM.

[25] Garreth W. Tigwell and David R. Flatla. 2016. Oh That's What You Meant!: Reducing Emoji Misunderstanding. In *Adjunct Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM.

[26] Leticia Vidal, Gastón Ares, and Sara R. Jaeger. 2016. Use of Emoticon and Emoji in Tweets for Food-related Emotional Expression. *Food Quality and Preference* 49 (2016), 119 – 128.

[27] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2016. EmojiNet: Building a Machine Readable Sense Inventory for Emoji. In *Social Informatics*. Springer International.

[28] Sanjaya Wijeratne, Lakshika Balasuriya, Amit P. Sheth, and Derek Doran. 2017. A Semantics-Based Measure of Emoji Similarity. In *Proceedings of the International Conference on Web Intelligence*.

[29] Sarah Wiseman and Sandy J. J. Gould. 2018. Repurposing Emoji for Personalised Communication: Why 🍕 Means "I Love You". In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

[30] Rui Zhou, Jasmine Hentschel, and Neha Kumar. 2017. Goodbye Text, Hello Emoji: Mobile Communication on WeChat in China. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

Hanmoji: What Chinese Characters and Emoji Reveal About Each Other

Jennifer 8. Lee
Emojination
San Francisco, California, USA
jenny@emojination.org

Jason Li
Visual Culture Studies, The Civic Beat
Hong Kong, Hong Kong
jason@thecivicbeat.com

An Xiao Mina
Visual Culture Studies, The Civic Beat
Hong Kong, Hong Kong
an@thecivicbeat.com

ABSTRACT

Chinese characters predate the introduction of digital emoji by approximately 3,000 years. Despite the temporal gap, there are striking parallels between the canonical set of Chinese radicals and the set of emoji that have currently been approved by the Unicode Consortium. Comparing the 214 Kangxi Chinese radicals with the 3,019 emojis in the Unicode 12.0 set can reveal semantic gaps and provide directions for new emoji. Our analysis found that 72.4% of radicals have reasonable emoji equivalents, while only 17.8% of radicals lack any emoji equivalent that we could determine.

CCS CONCEPTS

• **Applied computing** → **Language translation**; **Media arts**.

KEYWORDS

Emoji; Chinese; Chinese Radicals; Kanji; Kangxi; Linguistics

ACM Reference Format:
Jennifer 8. Lee, Jason Li, and An Xiao Mina. 2019. Hanmoji: What Chinese Characters and Emoji Reveal About Each Other. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316543

1 BACKGROUND

The earliest known forms of Chinese writing is found on oracle bones, dating to the Shang Dyanasty (18th–12th Century BC). Chinese scripts evolved through various stages: guwen ("ancient figures") found in inscriptions from the period around the late Shang dynasty (c. 1123 BC), the dazhuan ("Great Seal") scripts of the the Zhou dynasty (1046 to 256 BC), and the modern day forms, which were largely established in the Qin period (221–207 BC). Many of the original oracle writing were based on visual representations of natural objects, such as "horse", "dog", and "mouse".

The origin of modern digital emoji is now commonly cited to be 1999, when Japanese designer Shigetaka Kurita created what has become the first popular and widely used set, consisting of 176 emoji, for DoCoMo. The Unicode Consortium became the body tasked to unify the disparate sets from various carriers. Starting with the 625 emoji characters in the 2010 Unicode Standard, the set now numbers 3,019 emojis as of the Unicode 12.0 release, a count which includes the gender and skintone variations. Since emoji stem from Japanese culture, and aspects of Japanese writing, specifically the kanji characters, are derived from Chinese characters, there is a natural cultural link from Chinese characters to emoji.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316543

2 METHOD

We assembled the set of 214 Kangxi Chinese radicals which have become the de facto ordering standard for Chinese dictionaries since their original introduction in the 1,615 Zihui dictionary. We then aggregated the English definitions for the radicals from yellowbridge.com [1], Wikipedia, and Wiktionary, with a special effort to take a "broad" definition when the term was translated into English. While some concepts were direct and straightforward – sun, water, fire, pig – others were more conceptual, such as (mi4), which variously could be "over", "crown", "3-side enclosure from the top", "cloth cover".

We then went through the 214 radicals and tried to match the radicals with 1) direct emoji equivalents, where the definition is essentially equivalent, 2) indirect emoji equivalents, where the meaning can be understood even if the definition is not equivalent, and 3) associative emoji meaning, where if forced to find an emoji that represents an idea, we have one that contains it as a quality of a component. Figure 1 shows the first 55 Chinese radical characters we explored, while the full list can be accessed at http://bit.ly/hanmoji

We made a special categorization of radicals which represented visual stroke, without a specific meaning – such as (slash, radical 4) and 匚(fang1, radical 22). For those without direct or indirect emoji parallels, we then methodically asked whether an emoji could reasonably be added, either because it represented a notable semantic gap, and if such an image could reasonably be expected to pass the specific criteria laid out by the Unicode Consortium (http://unicode.org/emoji/proposals.html), which include whether a proposed emoji is visually distinct and is not transient (e.g. has demonstrated longevity).

3 RESULTS AND DISCUSSION

Of the 214 Kangxi radicals, we count that 84, or 39.3%, have reasonably direct emoji equivalents as of the March 2019 Unicode 12.0 release, which had 3,019 emojis, including all the gender and skintone variations [2]. In addition, an additional non-overlapping 71, or 33.2%, have reasonable indirect emoji equivalents for at least one of their meanings. And third, we found 21, or 9.8%, emoji that had associative qualities such as 🍯 (HONEY POT) for 甘(sweet). So 72.4% of radicals have reasonable equivalents, while only 17.8% of radicals lacked any of the three.

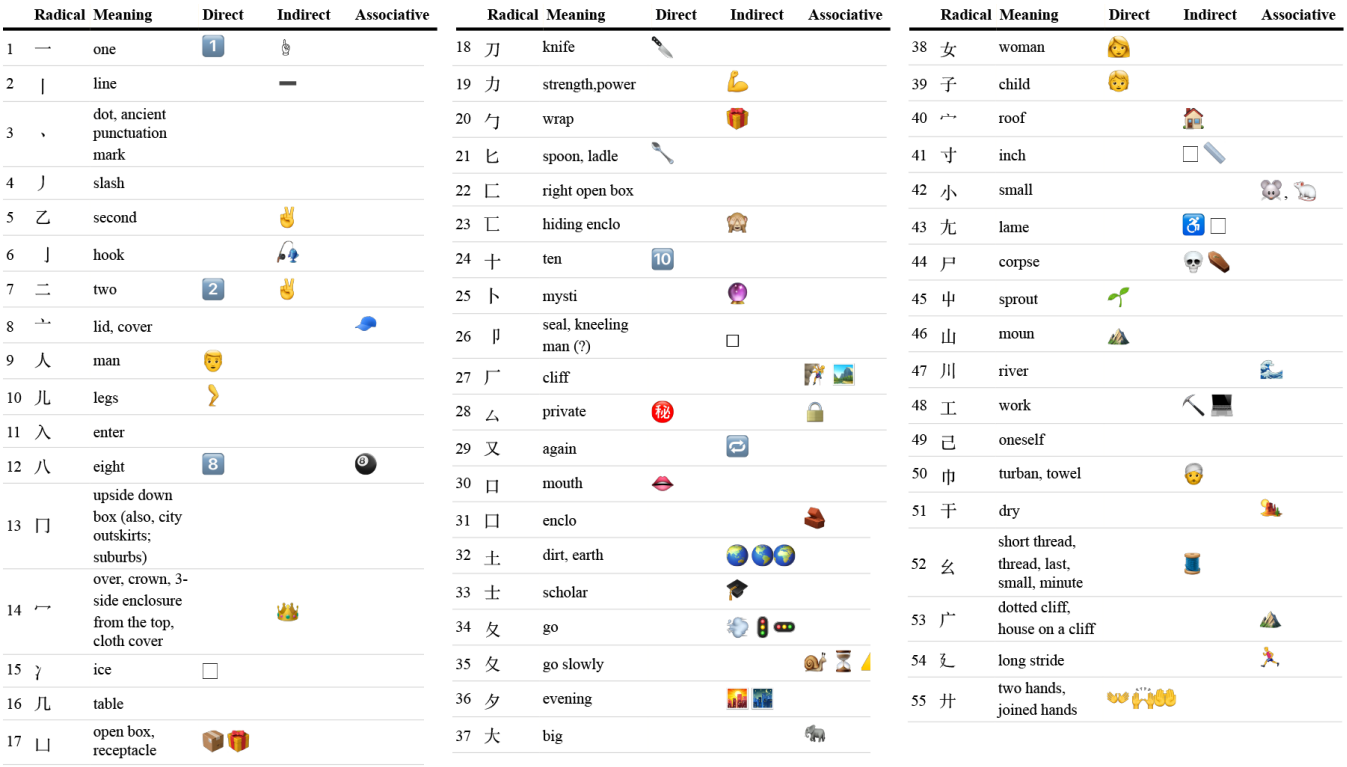


Figure 1: A comparison of the Chinese Kangxi Radicals and direct, indirect and associative emoji equivalents. Only the first 55 radicals are shown. The full list of the 214, along with an expanded chart, can be accessed at - http://bit.ly/hanmoji)

We also identified 14 Kangxi radicals, or 6.5%, that had reasonably strong chance of having either direct or indirect emoji equivalents based on the current Unicode selection factors, which include visually representable and distinct, longevity, significant semantic meaning, and popular demand. These include 龠(FLUTE), 豆(BEAN), 支(BRANCH), 穴(CAVE), 几(TABLE), and 矛(SPEAR). Next, we identified 15 Kangxi radicals, or 7.0%, that could possibly have emoji equivalents passed, but we categorize them as weaker than the previous set because of difficulty of visual representation or contemporary relevance. These include radicals such as 玉(jade), 戈(spear-axe), 臼(mortar), and 鼎(tripod).

Perhaps most interesting is that we identified a few clusters of other Kangxi radicals that fairly atomic and have important semantic meaning, but would face an uphill battle to have direct or indirect emoji equivalents passed because of the difficulty of visual representation. Some of them do have associative emoji meanings, however. These are:

(1) Adjectives: 大(big), 小(small), 長(long), 高(tall or high).
(2) Conveying "selfness" or privacy: 己(self), (private), 自(oneself).
(3) General concepts that may be too broad to tie to one distinct visual representation: 生(life), 用(to use), and 辛(bitter).
(4) Last we identified a cluster of radicals that have decent semantic meaning, but also face an uphill battle because of the difficult of representation; Textures: 皮(SKIN), 髟(HAIR), 革(LEATHER), 毛(FUR), (BRISTLE).

But the Chinese radicals are also interesting for which semantic areas are nearly completely absent, yet have significance presence in emoji, and often at an "atomic" level. For example, there are almost no emotions contained in the Kangxi Chinese radical set. In contrast, the most popular, emoji are almost always the yellow "smiley" faces which show concepts 😢 (Crying Face), 😠 (Angry Face), 😀 (Grinning Face).

4 CONCLUSION

Notably 170, or 79.4%, of the 214 Kangxi radicals have a direct or indirect emoji equivalent, or a reasonable chance of getting one. That speaks to the universality and longevity of certain concepts in a human existence, which include animals, body parts, colors, natural elements, foods, and tools. The analysis also helped identify certain semantic gaps, including basic adjectives, general concepts, and items which should give the Unicode Consortium a basic roadmap. These include 龠(FLUTE), 豆(BEAN), 支(BRANCH), 穴(CAVE), 几(TABLE), and 矛(SPEAR).

REFERENCES

[1] Yellow Bridge. 2019. List of Chinese Radicals. (2019). https://www.yellowbridge.com/chinese/radicals.php
[2] Unicode Consortium. 2019. Emoji List, v12.0. (2019). https://unicode.org/emoji/charts/emoji-list.html

Exploring (Dis-)Similarities in Emoji-Emotion Association on Twitter and Weibo

Mingyang Li
University of Pennsylvania
Philadelphia, PA, USA
myli@seas.upenn.edu

Sharath Chandra Guntuku
University of Pennsylvania
Philadelphia, PA, USA
sharathg@sas.upenn.edu

Vinit Jakhetiya
Indian Institute of Technology
Jammu, J&K, India
vinit.jakhetiya@iitjammu.ac.in

Lyle H. Ungar
University of Pennsylvania
Philadelphia, PA, USA
ungar@cis.upenn.edu

ABSTRACT

Emojis have gained widespread acceptance, globally and cross-culturally. However, Emoji use may also be nuanced due to differences across cultures, which can play a significant role in shaping emotional life. In this paper, we a) present a methodology to learn latent emotional components of Emojis, b) compare Emoji-Emotion associations across cultures, and c) discuss how they may reflect emotion expression in these platforms. Specifically, we learn vector space embeddings with more than 100 million posts from China (Sina Weibo) and the United States (Twitter), quantify the association of Emojis with 8 basic emotions, demonstrate correlation between visual cues and emotional valence, and discuss pairwise similarities between emotions. Our proposed Emoji-Emotion visualization pipeline for uncovering latent emotional components can potentially be used for downstream applications such as sentiment analysis and personalized text recommendations.

CCS CONCEPTS

• **Human-centered computing** → **Social networks**.

KEYWORDS

Emoji; Emotions; Weibo; Twitter; China; United States

ACM Reference Format:
Mingyang Li, Sharath Chandra Guntuku, Vinit Jakhetiya, and Lyle H. Ungar. 2019. Exploring (Dis-)Similarities in Emoji-Emotion Association on Twitter and Weibo. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316546

1 INTRODUCTION

Across countries, interpretations of Emojis may be nuanced due to cultural differences, which can play significant role in shaping emotional life [17]. Same Emojis can be differently associated with emotions to people from different cultures [39]. If quantitatively analyzed, such differences can inform research for cross-cultural linguists, personalized recommender systems for travelers, and downstream applications for creative designers who employ Emojis in their products. However, related quantitative research often focuses on usage statistics within individual countries, such as the *SwiftKey* Emoji Report [38], and cross-cultural comparison in Emoji-Emotion similarities is still a growing field of research.

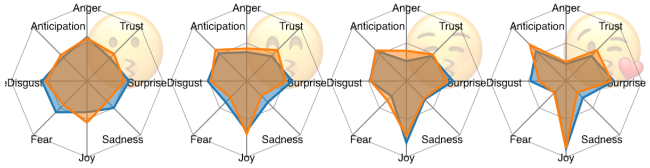


Figure 1: Associations of kissing face Emojis to 8 basic emotions. On each radar chart (orange: Sina Weibo; blue: Twitter), Center points, inner octagons, and outer octagons indicate −100%, 0, and 100% similarities, respectively. With more visual cues (curly eyes, blushes, and a heart mark) added to the Smiley Emoji, the emotional valence increases.

In this paper, we present a methodology to uncover latent emotional components associated with Emojis using more than 100 million posts from China (Sina Weibo) and the United States (Twitter). These platforms are the closest analogs in both countries considering the demographics and similarities in degrees of affective and informative content [11, 23, 25]. Starting with learning vectorial representations for tokens in each platform, we quantify similarities of Emojis with 8 basic emotions. Immediately, this enables us to infer correlation between visual cues and emotional valence, as demonstrated in Figure 1. Furthermore, Emoji-Emotion associations can be compared across distinctive languages, enabling more possibilities in cross-cultural research.

2 BACKGROUND AND RELATED WORK

Weibo and Twitter Platforms. Social media posts on Twitter and Weibo are predictive of several traits, including users' demographics [36, 45], personality [13, 21], location [35, 46], psychological stress [12, 22], and mental health [15, 40]. In this paper, we study latent emotional components in Emojis using Emoji-Emotion associations across the US, represented by Twitter posts, and in China by Weibo posts.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316546

Differences in Emotion Expressions and Perceptions. There are similarities as well as differences in how people of different cultural backgrounds express and interpret emotions. First, evolutionary and biological processes generate universal expressions and perceptions of emotions [1, 42]. For example, facial expressions is one of such universal channels that convey emotions across populations [8]. On the other hand, culture can play a significant role in shaping emotional life. Specifically, different cultures may value different types of emotions (e.g., Americans value excitement while Asians prefer calm) [41], and there are different emotional display rules across cultures [27]. In general, psychological research reveals both cultural similarities and differences in emotion expressions and perceptions [9]. Prior works also suggest that culture plays a key role in predicting perceptions of affect [14, 16, 47]

Lexical Approach to Measuring Emotions. Psychologists have long been employing curated lexicons in their researches [18]. For emotion analysis, the NRC Word-Emotion Association Lexicon (EmoLex, or simply NRC) is a popular dictionary consisting of 14,182 unigrams mapped to 10 'Affect Categories' [30]. Twitter is one of the platforms where EmoLex has been shown to be helpful in predicting sentiment analysis [6]. Prior work has also compared Twitter to Weibo using lexical approach in emotion usage [37].

However, cross-platform, and specifically cross-cultural, differences in Emoji-Emotion associations are less explored. Analyzing use of Emojis between the US and Chinese contexts provides an opportunity to assess a plethora of behaviors related to emotion expression and, arguably, emotional symbols that are culturally embedded.

Prior Studies on Emojis. Prior Emoji-related psychological research mainly focuses on (1) rendering-specific interpretations to Emojis, (2) Emoji-assisted sentiment annotation, and (3) embedding-derived insights of Emoji use. In the former topic, researchers systematically studied Emoji misinterpretation based upon rendering used by different operating systems (Android, iOS and Windows) [29]. Cross-platform mapping to correct the misinterpretation by applying it to the task of sentiment analysis was also proposed [31]. In sentiment-focused studies, a linguistic approach is often employed. For example, researchers studied causal inference to test whether adopting Emojis would cause individual Twitter users to employ fewer emoticons in their text [32] and also examined the effect of diversity of Emoji set in learning representations of emotional content in texts [10]. With the growing popularity and promise of word embedding algorithms, Emojis' semantic association with different word tokens using distributional semantics was carried out [2]. Meanwhile, there were also studies focused on how Unicode descriptions could indicate sentimental components in the Emojis [7]. Later, EmojiNet, the largest machine-readable Emoji sense inventory that links 2,389 Unicode Emoji representations to their English meanings extracted from the Web was released [44]. Researchers also studied the non-compositional nature of multi-Emoji expressions using frequency distribution of the words and sentiment analysis of the tweets containing them [24].

3 METHODS

3.1 Data Collection

This study received approval from authors' Institutional Review Board (IRB). We obtained Twitter data from a 10% archive released by the *TrendMiner* project [33], which exploited a streaming API from Twitter. Since Weibo lacks a streaming interface (as Twitter) for downloading random samples over time, we queried for all posts from individual users. The list of users were crawled using a breadth-first search strategy beginning with random users. On both platforms, we limited our analysis to posts created in the year of 2014 (to avoid potential confounds in the adoption of skin tones introduced in 2015). We obtained approximately 31.8 million posts on Twitter and 136 million posts on Weibo.

3.2 Pre-processing

Geo-location: On Twitter, the coordinates and tweet country location (whichever was available) was used to geo-locate tweets. On Weibo, user's self-identified profile location was used to identify the geo-location of messages.

Language Filtering: To remove the confounds of bilingualism, we filtered posts by the languages in which they were composed. Language used in each post is detected using a pre-trained *fastText* model released by Facebook [4, 19, 20]. Non-English tweets in the Twitter corpus and non-Mandarin posts on Weibo are removed. Particularly in the Weibo dataset, traditional Chinese characters were converted to Simplified Chinese using *hanziconv* Python package [1] to conform with the EmoLex lexicon used in later sections. We also remove any direct re-tweets (indicated by '`RT @USERNAME:`' on Twitter and '`@USERNAME//`' on Weibo).

Tokenizing: Twitter text was tokenized using *Social Tokenizer* bundled with *ekphrasis* [3], a text processing pipeline designed for social networks. Weibo posts were segmented using *Jieba*[2] considering its ability to discover new words and Internet slang, which is particularly important for a highly colloquial corpus like Sina Weibo. Using *ekphrasis*, URLs, email addresses, percentages, currency amounts, phone numbers, user names, emoticons and time-and-dates were normalized with meta-tokens such as '`<url>`', '`<email>`', '`<user>`' etc.

3.3 Training Embedding Models

To study the lexical semantics on both platforms, *fastText* models were trained on each corpus. These models were trained for 10 epochs with learning rates initialized at .025 and allowed to drop till 10^{-4}. The dimension of learned token vectors was chosen to be 100 based on previous work. To counter side effects due to the randomized initialization in the *fastText* algorithm, each model was trained for 5 times independently with identical parameters. The 10 models are referred to as $\{m_{t,1}, m_{t,2}, ..., m_{t,5}, m_{w,1}, m_{w,2}, ..., m_{w,5}\}$, where $c \in \{t, w\}$ refers to Twitter and Weibo, respectively. Each model can provide a vectorial representation to a token of interest.

To provide an intuitive view on the vectors obtained, we compressed the vectorial representations of Emojis from $m_{t,1}$ and

[1] https://pypi.org/project/hanziconv/
[2] https://github.com/fxsjy/jieba

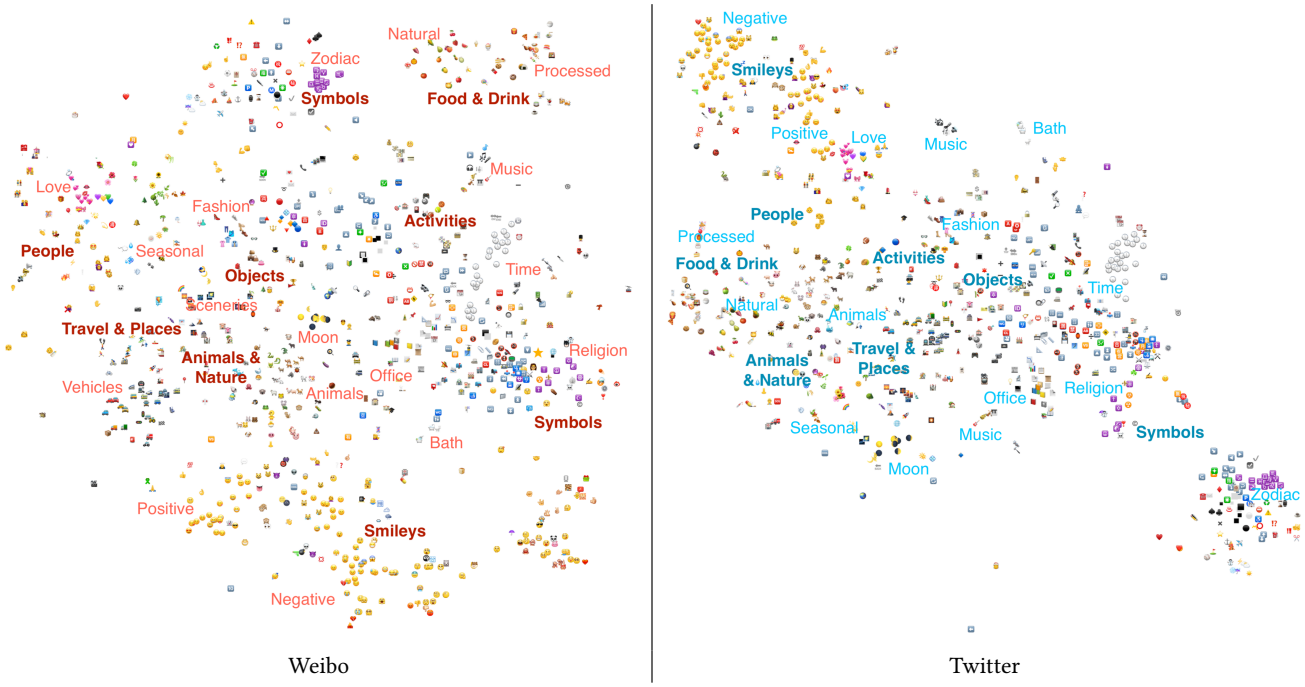


Figure 2: t-SNE visualization of the learned vectors for Emojis on both platforms. Weibo is shown on the left, and Twitter is put on the right. Textual labels are manually added afterwards for clarity. Dark labels are names of Unicode categories, while light-colored labels are manually created classifications that does not necessarily map to Unicode categories.

$m_{w,1}$ into two dimensions using the t-SNE algorithm [26]. The two-dimensional coordinates are plotted in Figure 2 side by side. Naturally-forming clusters are labeled in-place for clarity.

On both platforms, Emojis tend to cluster by semantic usage. Two patterns are worth noticing: 1) coherent to how bag-of-words model works [28], tokens with limited possible usages appear closer (e.g. clock faces, zodiac signs, and moon phases), while less topic-specific Emojis tend to spread away (e.g. smileys, objects, and symbols). 2) clusters resemble the categories of Emojis defined by the Unicode Consortium[3]. For example, the clusters formed by clock-face Emojis may be resulted from the common usage of the sentence structure, 'Let's meet at [clock-face Emoji here]'. Similarly, the cluster of transportation tools may be a result of many 'Let's go there by [transportation Emoji]' sentences. Despite that t-SNE visualizations can be misleading at times [43], these clearly separated clusters here suggest that our *fastText* training has been successful in capturing usages of Emojis.

3.4 Projecting Emojis onto Emotions

Within each model, $m_{c,i}$ where corpus $c \in \{t, w\}$ and instance $i = 1, ..., 5$, for each of the 8 basic emotions in the EmoLex, $T \in$ {Anger, Joy, ..., Trust}, all tokens associated with this emotion are averaged into a (model-specific) 'emotion vector' $\vec{T}_{c,i}$. Since all EmoLex tokens are verbal, which can render as a bias when their corresponding vectors are compared with those of the Emojis (which are non-verbal tokens), the average vector of all 8 emotions are then subtracted from each of the 8 emotion vectors, essentially zeroing out the center of the polyhedron defined by the 8 emotions. The average of all Emoji vectors is also subtracted from each of the Emojis for similar reason. For each pair of emotion ,$\vec{T}_{c,i}$, and Emoji, $\vec{J}_{c,i}$, in this model, their cosine similarity, denoted by $s_{J,T,c,i}$, is computed. The similarities between all Emojis and all emotions are then assembled into a 'similarity matrix' $\mathbf{s}_{c,i}$. The 5 similarity matrices across the 5 instances are then averaged for stability. The result is denoted by $\mathbf{s}_c$.

The similarity matrix $\mathbf{s}_c$ can be interpreted as follows. Each row of the matrix, $\vec{s}_{J,c}$, specifies the similarities between the Emoji J to each of the 8 emotions. Every column of the matrix, $\vec{s}_{T,c}$, contains the similarities between the emotion T to each of the Emojis. In other for the matrices to be comparable, we consider only the Emojis appeared in both platforms.

Data and source code to replicate the results in this paper are available at https://github.com/tslmy/EMOJI2019.

4 RESULTS AND DISCUSSION

4.1 How are Emojis associated with different Emotions?

For each Emoji J, its 2 similarity vectors (described in the previous section), $\vec{s}_{J,w}$ and $\vec{s}_{J,t}$, are plotted as 2 overlapping polygons. This radar chart, Figure 3, compares the (dis-)similarities between several Emojis (in every Unicode category) to each of the 8 basic emotions.

In the radar charts, center points, inner octagons, and outer octagons indicate −100%, 0, and 100% similarities, respectively. For

[3]https://unicode.org/emoji/charts/full-emoji-list.html

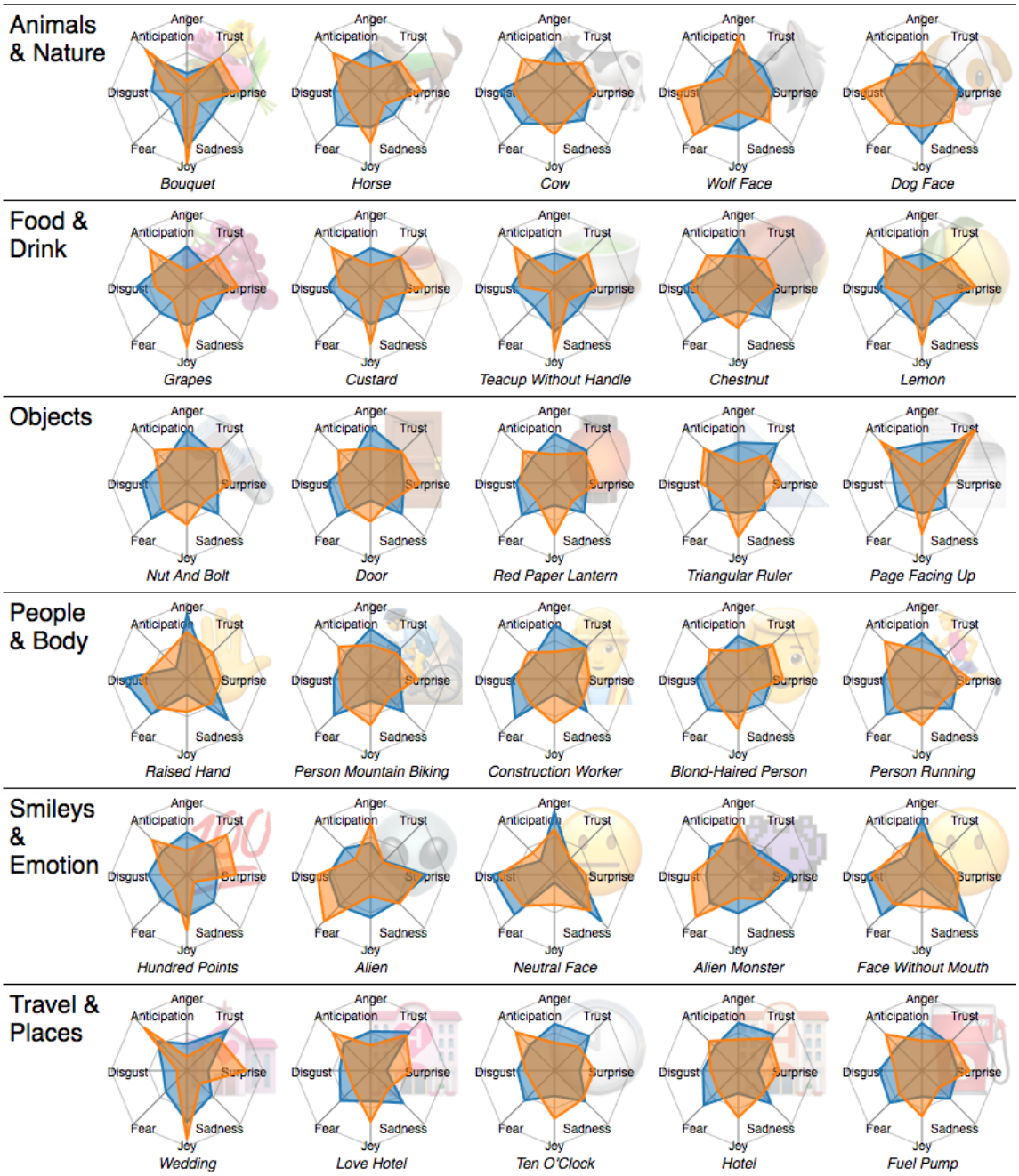


Figure 3: Radar chart comparing Emoji-Emotion similarities across the two platforms (orange: Weibo; blue: Twitter). From left to right, Emojis are sorted in increasing order of cross-cultural similarity (i.e., decreasing sum of absolute difference across platforms over the 8 emotions). Center points, inner octagons, and outer octagons indicate −100%, 0, and 100% similarities, respectively.

example, the Running Man Emoji 🏃 has near-zero similarities to all 8 emotions, indicating the potential emotionally neutral usage this Emoji.

Another counter-intuitively neutral Emoji is the Kissing Face Emoji 😗 featuring '3' shaped mouth and dotted eyes. It is interesting to compare this Emoji with the following Emojis: Kissing Face With Smiling Eyes 😙, Kissing Face With Closed Eyes 😚, and

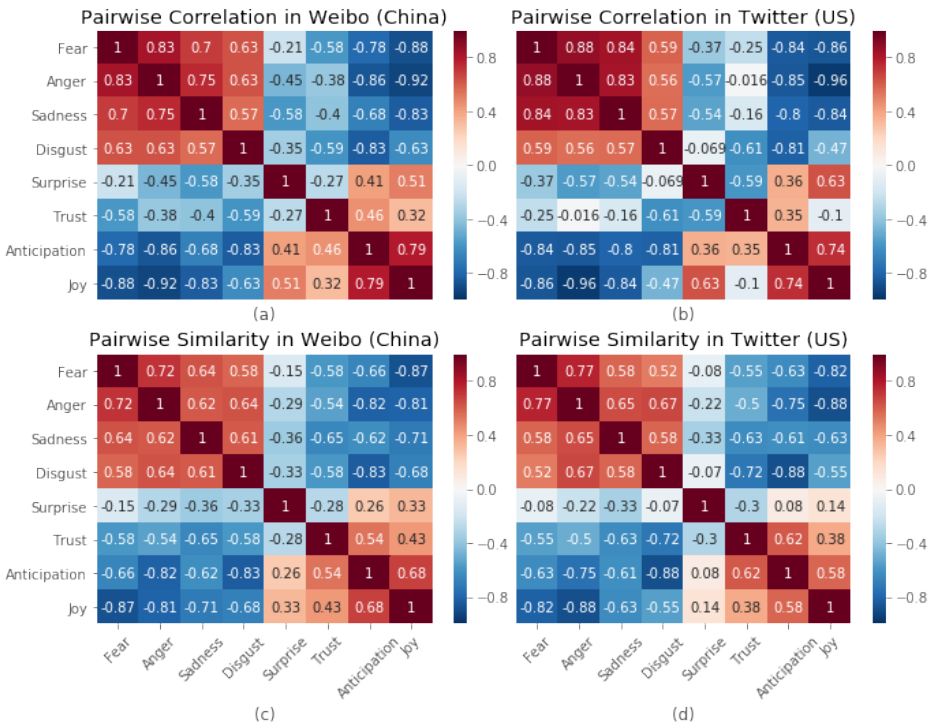


Figure 4: Pairwise correlations of emotions based on Emoji-Emotion cosine similarities in (a) Weibo and in (b) Twitter. Pairwise cosine similarities of emotions based on NRC emotion vectors in (c) Weibo and in (d) Twitter.

Face Blowing a Kiss 😘, as shown in Figure 1. With graphical features added, namely smiling eyes in the second, rosy cheeks in the third, and finally a heart in the forth, the radar charts demonstrate increasing positive valences in the emotions.

Besides comparing a set of Emojis with some similar features, the radar charts also enable us to compare emotional association of each Emoji across platforms, and potentially cultures. Each row in Figure 3 are Emojis segregated based on Unicode categories. Emojis that tend to be associated with negative feelings have large similarity values with Anger, Disgust, Fear, and Sadness, with a negative value in Joy and in Trust. Emojis categorized to be associated with positive feelings, on the other hand, usually have large positive value in Joy, negative values in Anger, Fear, and Sadness, and near-zero similarities with Anticipation, Surprise, and Disgust. This suggests recognizable normative and also culture specific patterns of Emoji-Emotion associations.

4.2 Pairwise Correlations of Emoji-Emotion Association

To further quantify the differences in Emoji-Emotion associations, we examined the pairwise correlations between Emoji-Emotion similarities on Weibo and Twitter. For each platform c, the Spearman correlation coefficient (SCCs) matrix of the 8 emotions' similarity vectors, $\vec{s}_{T,c}$, is computed. The two matrices are shown as heatmaps (Figure 4 (a) and (b)) to demonstrate pairwise correlations of emotions in each platform.

From the correlation matrix showing the average between Weibo and Twitter (Figure 4 (c) and (d)), we can observe that negative emotions, namely Fear, Anger, Sadness, and Disgust, are highly correlated with each other. Among the other 4 emotions, Anticipation and Joy are highly correlated, but Trust and Surprise are rather uncorrelated. While negative emotions seem to be universal, Surprise and Trust potentially have cultural nuances. This is particularly interesting when compared to prior research on the recognition scores of the corresponding facial expressions among Americans and Chinese [34].

The most variance correlated pair of emotions is Joy and Trust. While Joy is often (0.32) correlated with Trust in Weibo, they are inversely correlated (−0.10) in Twitter. This suggests that Weibo users treat trustworthiness as an enjoyable personal trait more than Twitter users do, coherent to a claim in [5].

Furthermore, the Surprise-Trust pair is more than twice as negatively correlated to Twitter users than to Weibo users. This may suggest that Twitter users expect more predictable behavior (fewer surprises) from trustworthy people than Weibo users do. Similarly, the heatmaps also suggest that Weibo users may not enjoy surprises as much as Twitter users do.

It is noted that that the correlations computed over Emoji-Emotion similarities are based on emotions expressed only in posts containing Emojis. We compare them pairwise cosine similarities between the 8 NRC emotion vectors. The pairwise similarity matrix between NRC emotions on Twitter and on Weibo are plotted in Figures 4

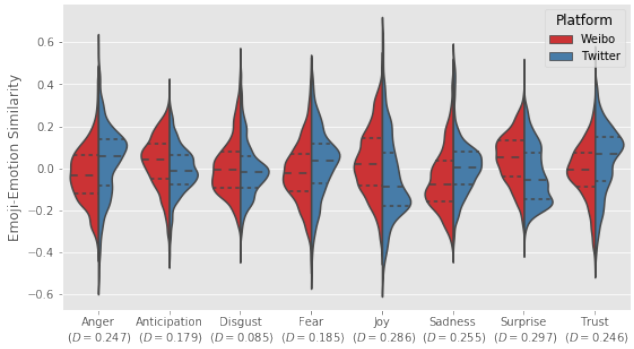


Figure 5: Violin chart showing distributions of Emoji-Emotion similarities for each emotion. Left halves and right halves show distributions on Weibo and on Twitter, respectively. Quartiles of the distributions are marked with dashed lines. A bandwidth of .2 is chosen to smooth out the distribution when plotting.

(c) and (d) respectively. Figures 4 (c) and (d) closely resembled Figures 4 (a) and (b), indicating that Emoji-Emotion associations also correlated with NRC Emotion similarities on both platforms.

4.3 Distributions of Emoji-Emotion Similarities for Each Emotion

Besides analyzing Emoji-Emotion similarities on a per-Emoji basis and by pairs of NRC emotions, an essential statistic is the distribution of these values. For each emotion T, its 2 similarity vectors, $\vec{s}_{T,w}$ and $\vec{s}_{T,t}$, are plotted as two sides of a violin chart, Figure 5. This chart compares how many Emojis the two platforms' users found similar to each of the 8 basic emotions.

An emotion that has a larger mean correlation to Emojis on Twitter than that on Weibo is more directly expressed by Twitter users, because this implies that Twitter users find many Emojis more relevant to the said emotion. In this sense, Twitter users express Anger, Fear, Sadness, and Trust more directly than Weibo users do, who more readily display Surprise, Joy, and Anticipation.

We report the Kolmogorov-Smirnov D statistic for each emotion in Figure 5. With a confidence level of 0.001 and identical numbers of Emojis considered (843), any emotion with a D-value exceeding 0.090 can be recognized as having distinct distributions of similarities to Emojis. All emotions except Disgust fit this criterion.

5 CONCLUSION

In this paper, we presented a method to learn latent emotional components of Emojis by learning vectorial representations of tokens on large-scale social media corpus from China (Weibo) and US (Twitter). The universality of the Unicode character set, to which modern Emojis belong, enabled us to compare Emojis across cultures, even if the selected languages share no common character.

Comparing latent emotional components of similar Emojis enables us to infer correlations between visual cues and emotion categories. As an example, we demonstrated in Figure 1 how the addition of visual features can increase the positive valence associated to kissing-face Emojis. In a similar visualization, we presented most differently used Emojis across platforms in Figure 3. Further, we studied pairwise correlations between emotions on each platform. The heatmaps (a) and (b) in Figure 4 implied that Twitter and Weibo users have different views towards Surprise and Trust, at least in terms of Emojis associated with them. The fact that (a) and (b) resembled the pairwise similarity matrices between emotions, (c) and (d), suggested that the correlations are a valid proxy for emotion analysis, strengthening our point. Finally, we studied the distributions of Emoji-Emotion similarities across platforms (Figure 5). We concluded that Twitter users and Weibo users indeed take different set of Emojis as suitable to each emotion with the exception of Disgust.

Though the current work is preliminary and insights are correlational in nature, future research can potentially infer insights into specific large scale psychological and cognitive phenomena by comparing the latent emotional contexts behind Emojis across Weibo and Twitter. Further, it would be interesting to study user behavior on Chinese users on Twitter and English users on Weibo.

REFERENCES

[1] Richard P Bagozzi, Nancy Wong, and Youjae Yi. 1999. The role of culture and gender in the relationship between positive and negative affect. *Cognition & Emotion* 13, 6 (1999), 641–672.

[2] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis.. In *LREC*.

[3] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 747–754.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* abs/1607.04606 (2016). arXiv:1607.04606 http://arxiv.org/abs/1607.04606

[5] Michael Harris Bond. 2010. *The Oxford handbook of Chinese psychology*. Oxford Library of Psychology.

[6] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2013. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2.

[7] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359* (2016).

[8] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[9] Hillary Anger Elfenbein and Nalini Ambady. 2003. Universals and cultural differences in recognizing emotions. *Current directions in psychological science* 12, 5 (2003), 159–164.

[10] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[11] Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. 2012. A comparative study of usersâĂŹ microblogging behavior on Sina Weibo and Twitter. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 88–101.

[12] Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes Eichstaedt, and Lyle Ungar. 2018. Understanding and Measuring Psychological Stress using Social Media. *arXiv preprint arXiv:1811.07430* (2018).

[13] Sharath Chandra Guntuku, Weisi Lin, Jordan Carpenter, Wee Keong Ng, Lyle H Ungar, and Daniel Preoţiuc-Pietro. 2017. Studying personality through the content of posted and liked images on Twitter. In *Proceedings of the 2017 ACM on web science conference*. ACM, 223–227.

[14] Sharath Chandra Guntuku, Weisi Lin, Michael James Scott, and Gheorghita Ghinea. 2015. Modelling the influence of personality and culture on affect and enjoyment in multimedia. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 236–242.

[15] Sharath Chandra Guntuku, J Russell Ramsay, Raina M Merchant, and Lyle H Ungar. 2017. Language of ADHD in adults on social media. *Journal of attention*

disorders (2017), 1087054717738083.

[16] Sharath Chandra Guntuku, Michael James Scott, Huan Yang, Gheorghita Ghinea, and Weisi Lin. 2015. The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–7.

[17] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 19 (2012), 7241–7244.

[18] Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality* 2, 3 (1988), 171–203.

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *CoRR* abs/1612.03651 (2016). arXiv:1612.03651 http://arxiv.org/abs/1612.03651

[20] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *CoRR* abs/1607.01759 (2016). arXiv:1607.01759 http://arxiv.org/abs/1607.01759

[21] Lin Li, Ang Li, Bibo Hao, Zengda Guan, and Tingshao Zhu. 2014. Predicting active users' personality based on micro-blogging behaviors. *PloS one* 9, 1 (2014), e84997.

[22] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 507–516.

[23] Xialing Lin, Kenneth A Lachlan, and Patric R Spence. 2016. Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on twitter and weibo. *Computers in Human Behavior* 65 (2016), 576–581.

[24] Rebeca Padilla López and Fabienne Cap. 2017. Did you ever read about Frogs drinking Coffee? Investigating the Compositionality of Multi-Emoji Expressions. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 113–117.

[25] Lin Ma. 2013. Electronic word-of-mouth on microblogs: A cross-cultural content analysis of Twitter and Weibo. *Intercultural Communication Studies* 22, 3 (2013).

[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html

[27] David Matsumoto. 1990. Cultural similarities and differences in display rules. *Motivation and emotion* 14, 3 (1990), 195–214.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[29] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. Blissfully happyâĂİ or âĂIJready to fightâĂİ: Varying Interpretations of Emoji. *Proceedings of ICWSM* 2016 (2016).

[30] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. 29, 3 (2013), 436–465.

[31] Fred Morstatter, Kai Shu, Suhang Wang, and Huan Liu. 2017. Cross-Platform Emoji Interpretation: Analysis, a Solution, and Applications. *arXiv preprint arXiv:1709.04969* (2017).

[32] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. Emojis on Twitter: A Causal Inference Approach. *CoRR* abs/1510.08480 (2015). arXiv:1510.08480 http://arxiv.org/abs/1510.08480

[33] Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[34] James A Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological bulletin* 115, 1 (1994), 102.

[35] Bahar Salehi, Dirk Hovy, Eduard Hovy, and Anders Søgaard. 2017. Huntsville, hospitals, and hockey teams: Names can reveal your location. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 116–121.

[36] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1146–1151.

[37] Kaisong Song, Shi Feng, Wei Gao, Daling Wang, Ge Yu, and Kam-Fai Wong. 2015. Personalized sentiment classification based on latent individuality of microblog users. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[38] SwiftKey. 2015. SwiftKey Emoji Report | The United States | Data. https://www.scribd.com/doc/262594751/SwiftKey-Emoji-Report

[39] Kohske Takahashi, Takanori Oishi, and Masaki Shimada. 2017. Is 😒 Smiling? Cross-Cultural Study on Recognition of EmoticonâĂŹs Emotion. *Journal of Cross-Cultural Psychology* 48, 10 (2017), 1578–1586. https://doi.org/10.1177/0022022117734372 arXiv:https://doi.org/10.1177/0022022117734372

[40] Xianyun Tian, Philip Batterham, Shuang Song, Xiaoxu Yao, and Guang Yu. 2018. Characterizing depression issues on Sina Weibo. *International journal of environmental research and public health* 15, 4 (2018), 764.

[41] Jeanne L Tsai, Brian Knutson, and Helene H Fung. 2006. Cultural variation in affect valuation. *Journal of personality and social psychology* 90, 2 (2006), 288.

[42] Yukiko Uchida and Shinobu Kitayama. 2009. Happiness and unhappiness in east and west: themes and variations. *Emotion* 9, 4 (2009), 441.

[43] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to use t-sne effectively. *Distill* 1, 10 (2016), e2.

[44] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. Emojinet: An open service and api for emoji sense discovery. In *Eleventh International AAAI Conference on Web and Social Media*.

[45] Wanru Zhang, Andrew Caines, Dimitrios Alikaniotis, and Paula Buttery. 2016. Predicting Author Age from Weibo Microblog Posts.. In *LREC*.

[46] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the eighth ACM international conference on web search and data mining*. ACM, 295–304.

[47] Yi Zhu, Sharath Chandra Guntuku, Weisi Lin, Gheorghita Ghinea, and Judith A Redi. 2018. Measuring Individual Video QoE: A Survey, and Proposal for Future Directions Using Social Media. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2s (2018), 30.

Emoji Prediction for Hebrew Political Domain

Chaya Liebeskind
Department of Computer Science, Jerusalem College of Technology, Lev Academic Center
Jerusalem, Israel
liebchaya@gmail.com

Shmuel Liebeskind
Department of Computer Science, Jerusalem College of Technology, Lev Academic Center
Jerusalem, Israel
israellieb@gmail.com

ABSTRACT

In this study, we aim to predict the most likely emoji given only a short text as an input. We extract a Hebrew political dataset of user comments for emoji prediction. Then, we investigate highly sparse n-grams representations as well as denser character n-grams representations for emoji classification. Since the comments in social media are usually short, we also investigate four dimension reduction methods, which associates similar words to similar vectorial representation. We demonstrate that the common Word Embedding dimension reduction method is not optimal. We also show that the character n-grams representations outperform all the other representation for the task of emoji prediction for Hebrew political domain.

CCS CONCEPTS

• **Information systems → Social tagging systems**; **Data analytics**; **Data mining**; **Content analysis and feature selection**; • **Computing methodologies → Machine learning approaches**; **Machine learning algorithms**.

KEYWORDS

Emoji prediction, Machine learning, Semantic analysis, Social media, Supervised learning.

ACM Reference Format:
Chaya Liebeskind and Shmuel Liebeskind. 2019. Emoji Prediction for Hebrew Political Domain. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3316548

1 INTRODUCTION

In recent years there has been increasing usage of emojis in social media. Emoji is defined as "a digital image that is added to a message in electronic communication in order to express a particular idea or feeling" (Cambridge dictionary). Any system that aims to address the task of modeling social media communication need to deal with the usage of emojis.

Following Novak et al. [36], Barbieri et al. [3] argued that efficient prediction of the most likely emoji given the short text of a message may help to improve different Natural Language Processing (NLP) tasks. These NLP tasks include both objective tasks, such as information retrieval, social media content generation and emoji suggestion, and subjective tasks like sentiment analysis and emotion recognition. They introduced the task of emoji prediction in Twitter and showed that their neural network model can outperform humans.

Inspired by the promising results of Barbieri et al. [3], Barbieri et al. [4] proposed the first shared task on multilingual emoji prediction. Previous findings about the idiosyncrasy of use of emojis across languages motivated Barbieri et al. [4] to focus on emoji prediction for two languages, English and Spanish. Additional researchers followed them and investigated the task for other languages including Italian [40], Japanese [42], Hindi, Bengali and Telugu [12, 13]. We are interested in emoji prediction for Hebrew. Hebrew is characterized by highly productive morphology and, as far as we know, has not been investigated before.

Since most of the Hebrew speakers live in Israel, we examined the statistics of social media usage in Israel (see Figure 1, generated by http://gs.statcounter.com/social-media-stats/all/israel). Facebook is leading with 77.45% of the users. Twitter is in the fourth place with only 3.99% of the users. Therefore, even though, most prior work used Twitter data, in order to gather a sufficient amount of data we used another social network, Facebook.

In this paper, we first describe emoji usage in our political Facebook dataset. Then, we explore the emoji prediction task. Given only a comment text, we aim to classify it to one of the twenty emojis that occur most frequently in the dataset. To perform this task, we adopt a supervised Machine Learning (ML) approach. We compare four semantic vector representations, i.e., Latent Semantic Analysis (LSA) [15, 25], Latent Dirichlet Allocation (LDA) [10, 22], Random projection (RP) [9] and Word Embedding [31, 32], to represent each comment as a vector of features. Each of the representations is generated by a different dimension reduction method, which associates similar words to similar vectorial representation. Moreover, all of the representations are built using entirely unsupervised distributional analysis of unlabeled text. We demonstrate that, in the case study of Hebrew Facebook, the Word Embedding dimensional reduction method achieves the lowest F1.

Additionally, we analyze two text representation approaches, i.e., n-grams and character n-grams. Both of these representations outperform all the semantic representations. We also show that the best character n-grams representation outperforms the FastText baseline [24], which is often on par with deep learning classifiers in terms of accuracy.

The rest of this paper is organized as follows: Section 2 introduces relevant background about emoji prediction. Section 3 presents the emoji prediction task, dataset, and the representation methods. Section 4 introduces the experimental setting, the experimental

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316548

results and a deep analysis of the best method. Finally, Section 5 summarizes the main findings and suggests future directions.

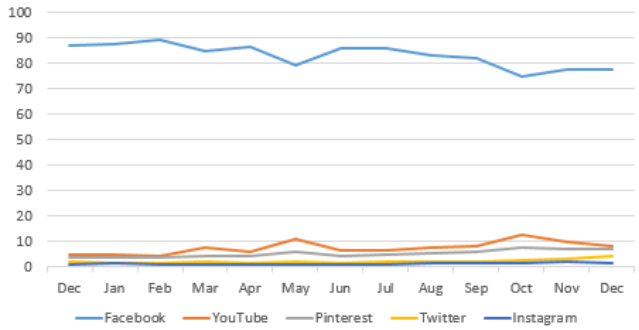


Figure 1: Social Media Stats Israel, Dec 2017 - Dec 2018

2 BACKGROUND

Recently, there is growing interest in distributional embeddings of emojis. Barbieri et al. [8] used the skip-gram neural embedding model [32] to train emoji embeddings from a large Twitter dataset of over 100 million English tweets. They evaluated these pre-trained emoji representations on two tasks: a pair similarity and relatedness, and clustering. They showed that the representations improve accuracy on both tasks. The same emoji representation methodology was adopted to compare the meaning and usage of emojis across two Spanish cities: Barcelona and Madrid [5], and across different languages [6].

A similar neural embedding model for emoji similarity was explored by Pohl et al. [38]. They claimed that search is a critical problem of emoji entry, and thus emoji keyboards need to be optimized for search. Being able to compute the level of similarity between two emojis allows to place related emojis close to each other and support users' search. They showed that the model obtains good performance in capturing detailed relationships between emojis.

Eisner et al. [16] argued that Barbieri et al. [8] method is unable to learn robust representations for infrequent emojis. They suggested estimating the representation of emojis directly from their description in the Unicode emoji standard. They found that their representation generally outperforms the emoji embeddings trained by Barbieri et al. [8] on the task of Twitter sentiment analysis.

To improve emoji embedding models, Wijeratne et al. [44] suggested incorporating more words by using longer emoji definitions. Using the information in EmojiNet [43], they considered three different ways to represent the meaning of an emoji, namely emoji descriptions, emoji sense labels, and the emoji sense definitions. They showed that their models outperform the previous best-performing emoji embedding model of Eisner et al. [16] on the sentiment analysis task. They also created a new publicly available dataset called EmoSim508 [43], which assigns human-annotated semantic similarity scores to a set of 508 carefully selected emoji pairs.

Currently, neural models have been used to model the semantics of emojis. Most of these models are based on neural networks with word embeddings representations which are generated by word2vec models [32] or gradient descent based learning algorithm.

Xie et al. [47] investigated the task of emoji recommendation in multi-turn dialogue systems. They proposed a hierarchical Long Short-Term Memory (LSTM) model to construct better dialogue representations by encoding the contextual information located in conversations. Their method significantly outperforms other LSTM models and a baseline bag-of-words model which takes logistic regression as the classifier and tf-idf as features.

Barbieri et al. [3] employed state-of-the-art Recurrent Neural Network (RNN) for emoji prediction. Their model is based on Bidirectional Long Short-term Memory Networks (BLSTMs) [18, 21] with two types of embedding representations: word embeddings and character-based continuous-space vector embedding [30]. They showed that the BLSTMs outperform a bag-of-words baseline, a baseline based on semantic vectors, and human annotators in this task.

Another variant of LSTM Networks [21, 41] was utilized by Felbo et al. [17] to predict sentiment, emotions and irony using an emoji prediction classifier as pre-trained system. Their pretrained model includes an attention mechanisms [48] to increase its sensitivity to individual words during prediction. Their model obtained state-of-the-art performance on 8 benchmark datasets. Based on Felbo et al. [17] architecture, Barbieri et al. [2] proposed a label-wise attention mechanism that is suitable for underrepresented emojis. They observed a performance improvement over Felbo et al. [17] model and over the efficient FastText [24] classifier.

By comparing emoji embeddings trained on a corpus of different seasons, Barbieri et al. [7] showed that some emojis are used differently depending on the time of the year . They proposed a method that considers temporal information for emoji prediction systems. The method extracts two types of embeddings: character BLSTM [30] and word embeddings. Then it concatenates the two representations (as in [3]) and passes them to a word LSTM and word attention units. They showed that their method outperforms state-of-the-art systems.

Guibon et al. [20] developed an automatic recommendation system based on user message analysis and real emoji usage. They showed that a multi-label RandomForest classifier with a bag-of-words/characters representation and computed features outperforms [3] BLSTMs.

Wu et al. [46] also addressed the multi-label setting of the emoji prediction task. They proposed a hierarchical neural model with attention mechanism. The model contains three modules, a character encoder to learn hidden representations of words using Convolutional Neural Networks (CNNs), a word encoder to learn sentence representations using a combination of CNN and LSTM, an emoji classifier to predict emojis for tweets. Their approach outperforms several baselines, including K independent bag-of-word Support Vector Machine (SVM) models, CNN as the word encoder, and a hierarchical model with LSTM in both character and word encoders, as well as humans in this task.

Recently, Barbieri et al. [4] introduced a new SemEval task, i.e., the multilingual emoji prediction task. The task was divided into two subtasks respectively dealing with the prediction of the emoji associated to English and Spanish tweets. For each subtask the

tweets that include one of the twenty emojis that occur most frequently in the Twitter data were selected. Therefore, the task can be viewed as a multi-label classification problem with twenty labels. The task required to predict the emoji by relying exclusively on the textual content of the text message. In total, 49 teams participated in the English subtask and 22 teams submitted a system run to the Spanish subtask. Even though many of the participating teams preferred neural architectures, mostly LSTMs and CNNs, the best performing system [14] on both English and Spanish datasets used a SVM classifier with bag-of-n-grams features (both characters and words). The task organizers suggested to extend the problem of modeling emoji semantics by incorporating more and more diverse languages.

One step in this direction was taken by Ronzano et al. [40]. They proposed the shared task also for the Italian language (ITAMoji) in the context of the Evalita 2018 evaluation campaign [11]. Twelve runs were submitted at ITAmoji by five teams. Generally, systems which employ neural network architecture obtained good performances in this task, especially when relying on BLSTM model.

The work of Tomihira et al. [42] was the first attempt to focus on Japanese. They collected Japanese tweets from Twitter and proposed a new model that learns from sentences. They investigated multiple models based on Encoder-Decoder model of RNN and CNN. In contrast to Zhao and Zeng [49] who showed that the CNN model have higher classification accuracy than the RNN model, they showed that in their case study the Encoder-Decoder model with attention was higher than the CNN model in accuracy and F1 score.

Another step in this direction was the study by Choudhary et al. [13] who created corpus for multilingual sentiment analysis and emoji prediction in Hindi, Bengali and Telugu. They addressed resource-poor languages because such discourse is available on resource-rich languages like English and Spanish, whereas resource-poor languages are largely ignored. Choudhary et al. [12] introduced a twin BLSTM RNNs model to learn emoji-based representations of resource-poor languages. They jointly train the resource-poor languages (Hindi and Telugu) with resource-rich languages (English and Spanish) in a common emoji space by using a similarity metric based on the emojis present in sentences from both languages. Their model outperforms the state-of-the-art emoji prediction approaches based on distributional semantics, semantic rules, lexicon lists and deep neural network representations without shared parameters. In this research, we also investigate emoji prediction in a resource-poor language, Hebrew.

3 EMOJI PREDICTION

3.1 Task

Emoji prediction is an anticipation task of predicting the emojis that appear in a given text message by relying exclusively on the textual content of that message.

Practically, we remove the emojis from the text of the messages and use them as labels. Therefore, the task can be viewed as a multi-label classification problem.

Following previous works on emoji prediction [3, 3, 40], we selected only messages with a single emoji, so that the challenge can be cast as a single-label classification problem, detailed examples from our Hebrew dataset are shown in Table 1[1].

Table 1: Examples from our Hebrew dataset

#	Comment	Label
1	איזו הפתעה יוזמה מבורכת aizw hpt`h iwzmh mbwrkt what a surprise, a blessed initiative	👏
2	שאפו כל הכבוד šapw. kl hkbwd well done! hats off to	👍
3	שבת מבורכת אשת חייל šbt mbwrkt ašt xiil a blessed Sabbath (Saturday) woman of valor	🌹
4	דייייי נקרע הלב diiii nqr` hlb enough, the heart is rending	💔
5	יהי זכרו ברוך ihi zkrw brwk of blessed memory	😢

3.2 Dataset

We have created a dataset for the task of emoji prediction using a political dataset by Liebeskind et al. [27–29]. All posts of Members of Knesset (MKs) between 2014-2016 (n=130 MKs, m=33,537 posts) have been downloaded via Facebook Graph API . The data included also the comments to these posts (n=5.37M comments posted by 702,396 commentators).

We have analyzed the emoji usage of the commentators. There are 786 types of emojis and 98,865 of the comments include at least one of them. There are 50,243 comments with a single emoji.

As mentioned in the previous section, only Facebook comments with a single emoji should be included in the task dataset. However, to increase the number of comments, we include comments with multiple appearances of the same emoji, such as חג שמח יא מלך 👑👑👑 (xg šmx ia mlk - happy holiday, yo king). After limiting our dataset to comments that contain one and only one **type** of emojis, our dataset contains 78,147 comments that include 593 emoji types.

Further analysis of the data reveals that there are 240 types of emojis which never appear alone. They either have multiple appearances or appear with additional types of emojis. The ten most frequent emojis that never appear alone with their frequencies and top-5 co-occurring emojis are presented in Table 2. In example #2 the emoji and all its co-occurring emojis are of the same *event* group[2]. In examples #6-#7, #9, and #10, the emoji's group (*geometric*, *transport-ground*, and *family*, respectively) is dominant in its top co-occurring emojis. In examples #3 and #4, two Japanese's buttons co-occur in all comments. In example #1, 🎤 belongs to the *music* group and its co-occurring emoji 🎹 belongs to the *musical-instrument*

[1]To facilitate readability, we used a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexico-graphic order, are *abgdhwzxTiklmns `pcqr št.*
[2]The group names are from the Full Emoji List, v11.0 of Unicode Emoji: https://unicode.org/emoji/

group. Since music is often played at wedding celebrations, it is also understandable that 🎤 and 💒 often co-occur in the same comment. In examples #5 and #8, it is difficult to interpret the connection between the emoji and its top co-occurring emojis, the connection can probably be explained by the comments' text. It is interesting to mention that in example #8 most of the top co-occurring emojis are from the *plant-flower* group (🌹(16), 🌷(14), 🌺(13), 🌸(13), 🌼(12), 🌻(12)).

Table 2: The 10 most frequent emojis that never appear alone with their frequencies and top-5 co-occurring emojis

#	Emoji	Frequency	Co-occurring emojis
1	🎤	40	😇(25) 💒(25) 👭(24) 🎹(24) 🍇(24)
2	🌄	37	🌆(23) 🎉(15) 🎈(9) 🎊(8) 🎁(6)
3	🈲	36	😇(35) 💒(35) 👭(34) 🉑(33) 🌹(30)
4	🉑	35	😇(35) 💒(35) 👭(34) 🈲(33) 🌹(29)
5	🌍	30	🏏(29) 🙌(28) 💩(28) 🍼(28) 👑(28)
6	⚪	27	🔵(17) ❤️(6) ⚫(3) 🔴(2) 💚(2)
7	🔵	27	⚪(17) ❤️(4) 🔴(3) 🔯(2) 👍(2)
8	💫	25	🌹(16) 😇(14) 🌷(14) ❤️(14) 🌺(13)
9	🚲	21	🚓(21) 🚑(21) 🚂(20) 🙌(20) 🌍(20)
10	👭	20	👭(13) 👭(8) 👰(7) 👴(4) 👍(4)

Emojis are used by 41,789 of the commentators. In this analysis that is focused on commentators we take under consideration only the top-100 commentators with high usage of emojis (above 50 comments) in our dataset. In Figure 2, we compare the average number of words in comments with emojis (the blue line) with the average number of words in comments without emojis (the green line). The x-axis' values denote the commentators, while the y-axis' values denote the average number of comments. Most of the commentators use a similar number of words with and without emojis. This implies that emojis are not used as a substitute for words. The second comparison in Figure 2 between the average number of words in comments with emojis (the blue line) and the average number of emojis (the orange line) also shows that more emojis do not mean less words.

Figure 3 shows the diversity of emojis in the comments. The x-axis' values denote the commentators, while the y-axis' values denote the number of comments (blue bars) or the number of emoji types (orange bars). The figure illustrates that the diversity of the emojis depends on user preferences rather than on the number of comments.

For completeness, we ranked the list of all the 702,396 commentators of our dataset by the number of their comments. The distribution of the top-100 commentators with high usage of emojis over this ranked list is presented in Figure 4. 6% of the emoji users are part of the top-100 heavy users of our dataset. Most of the emoji users are ranked between 1,000 to 10,000 (60%). Non of them is ranked below 15,583, all the top-100 commentators with high usage of emojis are part of the top-15,583 heavy users of our dataset.

For the task of emoji prediction, we selected the comments that include one of the twenty emojis that occur most frequently in the dataset we have described (see Table 3). After removing fake comments with more than 15 emojis, our dataset contains 45,410 comments.

Table 3: The 20 most frequent emojis with their comment frequency, total frequency, and average comment length

#	Emoji	Comment Freq	Total Freq	Avg Comment Len
1	👍	12,537	25,486	9.21
2	❤️	5,767	10,234	15.37
3	👏	4,206	14,572	7.73
4	🙏	3,728	6,359	16.62
5	😂	3,108	9,553	10.77
6	😊	2,624	3,216	15.74
7	❤️	1,880	3,639	14.98
8	🌹	1,787	3,254	11.93
9	😉	1,618	1,835	17.94
10	😍	1,453	2,662	8.48
11	👎	1,333	3,581	13.5
12	😉	1,324	2,016	15.12
13	👑	1,280	2,949	11.05
14	😊	1,222	1,532	20.2
15	💙	1,195	2,420	9.88
16	😀	1,189	1,938	14.12
17	😄	1,146	1,799	29.73
18	😘	1,114	1,886	11.89
19	✌️	995	1,979	11.43
20	😢	961	1,606	16.15

3.3 Method

In this research, we adopt a supervised Machine Learning (ML) approach for emoji prediction. The first step in training a classifier is deciding what features of the text are relevant, and how to encode these features. First, We investigate two types of text representations:

(1) N-grams representation - An n-gram is a contiguous sequence of n words. Each of the n-grams in the comment is considered as a feature. The score of the feature is the n-gram tf-idf
The n-grams representation is a high-dimensional sparse representation for documents of any length. However, the sparsity problem is much more critical for short texts, such as comments, where most words have only one occurrence.

(2) Character n-grams representation - Character n-grams are strings of length n. For example, the character 3-grams of the string "prediction" would be: "pre", "red", "edi", "dic", "ict", "cti", "tio" and "ion". Each of the character n-grams of the comment is considered as a feature and scored by its tf-idf. Since there is much less character combinations than n-gram combinations, character n-grams representation overcomes the problem of sparse data that arises when using n-grams

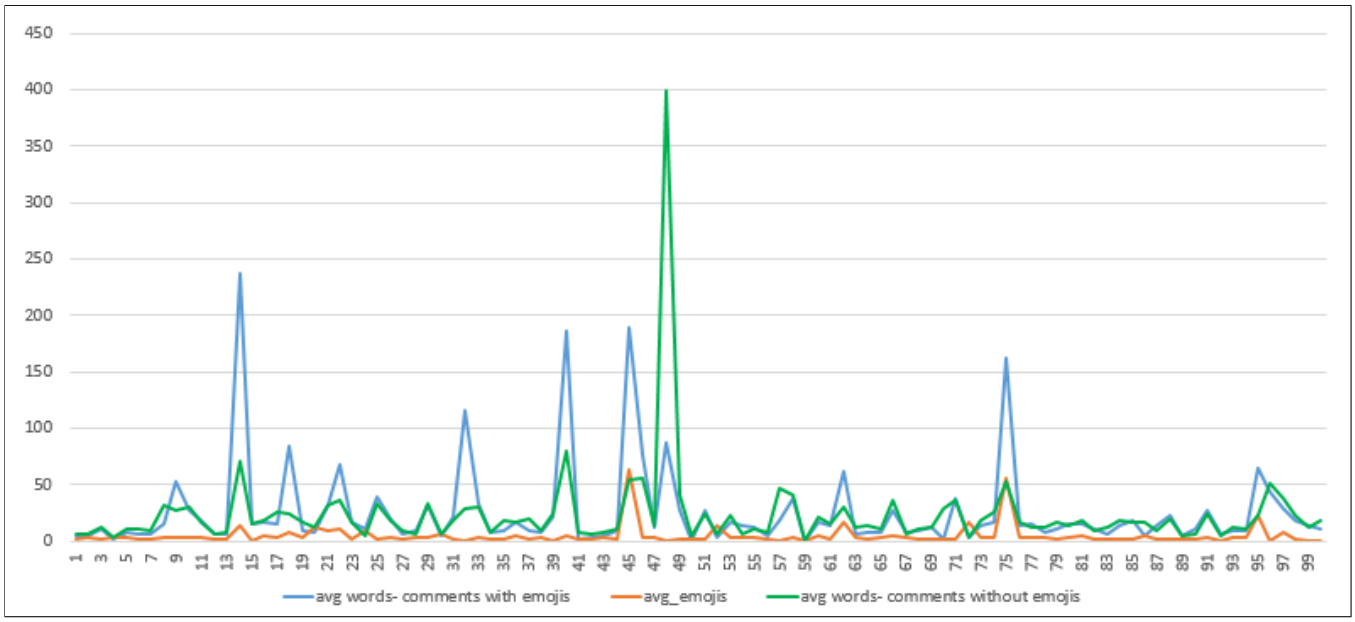


Figure 2: A comparison between the average number of words in comments with and without emojis

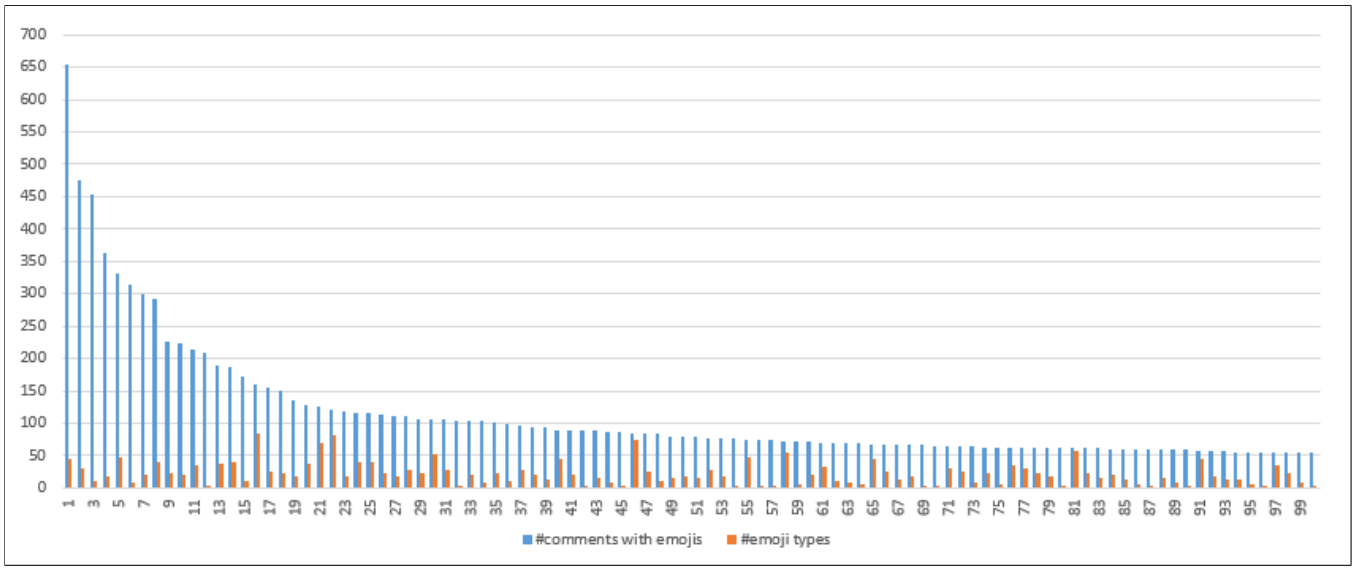


Figure 3: The diversity of emojis in the comments

representation. However, it still produces a considerably larger feature set.

Due to the tendency of noise and misspellings to have smaller impact on substring patterns than on n-gram patterns, character n-gram features can be quite effective for short informal text classification [1, 39].

We did not lemmatize the comments because previous work on Hebrew short text classification [29, 35] has indicated that lemmatization (using a Part-of-speech tagger) do not improve the performance.

Another approach to overcome the sparsity problem of the n-grams representation is to apply dimensional reduction methods for semantic analysis.

Next, we examine four semantic vector representations. Each of the representations is generated by a different dimension reduction method, which produces condensed vectorial representation of words in which similar words are described by similar vectors (with respect for instance to their cosine similarity). All of the representations are built using entirely unsupervised distributional analysis of large amount of unlabeled text.

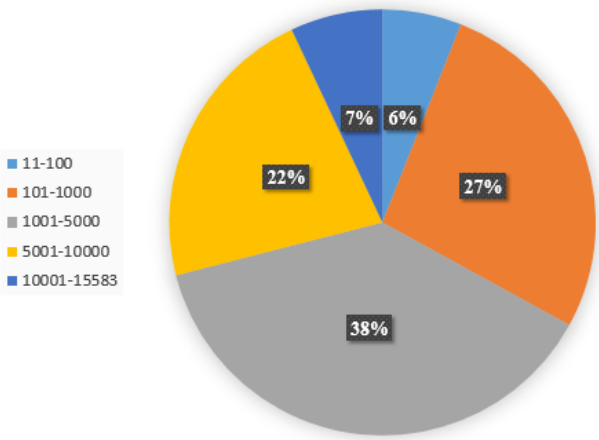


Figure 4: The distribution of the top-100 commentators with high usage of emojis over the ranked list all the dataset commentators

(1) **Latent Semantic Analysis (LSA)**[15, 25] constructs a semantic space from a large matrix of term-document association data by Singular Value Decomposition (SVD). SVD is a linear algebra procedure of decomposing an arbitrary matrix into three matrices, two of which are orthonormal and the third is a diagonal matrix whose diagonal values are the singular values of the matrix.

(2) **Latent Dirichlet Allocation (LDA)**[10, 22] is a generative statistical model for detecting latent semantic topics in large corpora. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all the words. The topic probabilities provide an explicit dense representation of a document. The main challenge of building an LDA model is how to estimate the distribution information of latent topics within the document. Different algorithms, such as Expectation Maximization (EM)[34] and Gibbs sampling [19] are used to face this challenge.

(3) **Random Projection (RP)**[9] projects the original high-dimensional data onto a lower-dimensional subspace using a random matrix whose columns have unit lengths. The main idea behind random projection is given in the Johnson-Lindenstrauss lemma [23]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved.

(4) **Word Embedding** [31, 32] approaches reduce the high dimensionality of words using neural probabilistic language models. A d-dimensional vector of real numbers models the contexts of each word. The vectors are meaningless on their own, but semantically similar words have similar vectors. Inspired by the methods for learning word vectors using neural networks, documents are also mapped to vectors (doc2vec) [26]. Given many contexts sampled from the document, the document vectors predict the next word in the given context. We use doc2vec to generate the word embedding based vectorial representation of the Facebook comments.

As far as we know, only the last semantic vector representation, Word Embedding, has been investigated for the task of emoji prediction, as detailed in Section 2.

4 EVALUATION

4.1 Evaluation Setting

While the supervised classification was performed on the comments that include one of the twenty emojis that occur most frequently in the Facebook data we collected (see Section 3.2), the unsupervised learning of dense vector representations was performed on the large corpus of 5.37M comments.

Table 4: The results of eight ML methods for all the four semantic vector representations

Rep.	Alg.	Precision	Recall	F1	Accuracy
RP	NB	0.1522	0.162	0.144	0.2045
	DT	0.1212	0.1228	0.1219	0.1851
	LR	0.233	0.1274	0.1302	**0.3065**
	RF	0.1514	0.1256	0.1277	0.2299
	MLP	0.1724	0.1652	**0.166**	0.2641
	SVC	0.443	0.0556	0.0298	0.2341
	AdaBoost	0.1186	0.0812	0.0655	0.2584
	Bagging	0.1559	0.1291	0.1316	0.2344
LDA	NB	0.1383	0.1016	0.0849	0.2536
	DT	0.1355	0.1313	0.1323	0.2174
	LR	0.1606	0.0919	0.077	0.2833
	RF	0.1618	0.1364	0.1401	0.2487
	MLP	0.2016	0.1308	0.1293	**0.2934**
	SVC	0.0114	0.05	0.0186	0.2291
	AdaBoost	0.1442	0.0901	0.0766	0.272
	Bagging	0.1691	0.1377	**0.1421**	0.2569
LSA	NB	0.156	0.1758	0.1464	0.1968
	DT	0.1472	0.1484	0.1477	0.2139
	LR	0.2255	0.1113	0.1139	**0.2712**
	RF	0.1726	0.1538	0.1559	0.2516
	MLP	0.1755	0.1449	0.1483	0.2622
	SVC	0.3429	0.0538	0.0261	0.2328
	AdaBoost	0.1629	0.1047	0.0967	0.2668
	Bagging	0.1767	0.1573	**0.1596**	0.2566
doc2vec	NB	0.1119	0.113	0.0911	0.1447
	DT	0.0865	0.0862	0.0863	0.1585
	LR	0.2522	0.0926	0.0797	0.3029
	RF	0.1114	0.1007	0.0971	0.2288
	MLP	0.1699	0.115	**0.1107**	**0.3061**
	SVC	0.1297	0.0707	0.0457	0.2784
	AdaBoost	0.1932	0.087	0.067	0.2901
	Bagging	0.1132	0.1016	0.0986	0.2305

Table 5: LogisticRegression and RandomForest results for the N-grams and Character n-grams representations

Representation	Alg.	Precision	Recall	F1	Accuracy
Char 2-grams	LR	0.3622	0.1664	0.1752	0.3718
	RF	0.2121	0.1713	0.1771	0.3271
Char 3-grams	LR	0.4365	0.1984	**0.2176**	**0.3847**
	RF	0.249	0.2072	0.2128	0.3412
Char 4-grams	LR	0.4588	0.191	0.2109	0.3792
	RF	0.2558	0.2069	0.215	0.3407
Char 5-grams	LR	0.4492	0.1855	0.204	0.3752
	RF	0.2482	0.2025	0.2112	0.336
Unigrams	LR	0.4284	0.184	0.2022	0.3712
	RF	0.2398	0.1943	0.2036	0.3279
Bigrams	LR	0.4335	0.1359	0.1509	0.3111
	RF	0.2106	0.1508	0.1645	0.2759
Trigrams	LR	0.3963	0.1047	0.1114	0.282
	RF	0.2484	0.113	0.1233	0.269

For classification, Scikit-learn[3] machine learning python module [37] was used. We used 5-fold cross-validation to estimate the classification performance, which is a resampling procedure of partitioning the dataset into a training set to train the model, and a test set to evaluate it. For dimensionality reduction, GenSim[4] python library with the default settings of 300 dimensions for the LSA, RP ,and Word Embedding (doc2vec) and 100 dimensions for LSA was used (due to computational limitations).

In our experiments, we compared the performance of our algorithms by four commonly used classification measures: precision, recall, F1, and accuracy. The scores are *macro-averaged*; we first calculate the measure for each label/emoji and then take the average of these scores.

4.2 Results

In our experiments, we combined the features in a supervised classification framework using eight ML methods: Bernulli Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Multilayered Perceptron (MLP), Support Vector Classification (SVC), Adaboost, and Bagging. Table 4 presents the results of all the dimensional reduction methods, extracted by the unlabeled data. When using the RP representation, MLP and Logistic Regression achieve the best F1 and accuracy, respectively. When using the LDA representation, Bagging and MLP achieve the best F1 and accuracy, respectively. When using the LSA representation, Bagging and Logistic Regression achieve the best F1 and accuracy, respectively. When using the doc2vec representation, the best F1 and accuracy are achieved by the same method, the MLP.

As detailed in Section 2, in previous works, the Word Embedding approach was exclusively selected as a dimensional reduction method. However, our results show that all the other reduction methods that we have suggested achieve a higher F1 score. The F1 advantage of the LDA representation, which is lower than the F1 of the RP and LDA representations, is statistically significant[5] at the 0.05 level.

The text representations generated a large feature set. Therefore, we filtered out features that have less than 8 appearances or appear in more than 5% of the comments in our dataset. We empirically selected two classifiers; LR and RF (out of the eight) that achieve good performance with reasonable run time.

Table 5 presents the results of the two types of text representation: N-grams representations (unigram, bigram, and trigram) and Character n-grams (character 2-grams, 3-grams, 4-grams, and 5-grams).

Generally, the LR classifier outperforms the RF classifier with all the text representation. The unigram representation and the

Table 6: The precision, recall, and F1 measure of the character 3-grams representation for each emoji

#	Emoji	Precision	Recall	F1
1	👍	0.33	0.84	0.47
2	❤️	0.62	0.69	0.65
3	🙏	0.45	0.44	0.44
4	👏	0.3	0.1	0.15
5	😂	0.36	0.39	0.38
6	😊	0.18	0.06	0.09
7	❤️	0.67	0.11	0.2
8	🌹	0.26	0.13	0.17
9	😉	0.16	0.04	0.06
10	😍	0.28	0.02	0.03
11	😜	0.17	0.02	0.04
12	😊	0.81	0.06	0.11
13	👑	0.58	0.34	0.43
14	😀	0.33	0.01	0.01
15	😄	0.95	0.11	0.2
16	😘	0.37	0.01	0.02
17	💙	0.53	0.24	0.33
18	👎	0.38	0.09	0.14
19	✌️	0.35	0.06	0.1
20	😢	0.64	0.22	0.33

[3]https://scikit-learn.org/stable/index.html
[4]https://radimrehurek.com/gensim/index.html
[5]In all the reported experiments, statistical significant was measured according to the two-sided Wilcoxon signed-rank test [45]

Table 7: Confusion matrix of the character 3-grams representation

	😘	😀	😜	😉	❤️	😢	🌹	😊	😄	👍	👑	👏	👎	😊	🙏	😂	✌️	💙	😍	❤️
😘	10	0	1	9	3	0	39	4	0	694	17	23	5	24	48	61	3	3	9	101
😀	0	6	7	23	4	2	46	0	0	762	4	29	7	35	41	127	3	5	0	57
😜	0	2	28	25	1	3	7	0	0	848	5	28	16	26	24	263	0	3	3	17
😉	1	0	21	65	0	1	7	0	0	1131	3	28	8	47	35	218	1	2	2	36
❤️	0	0	1	9	201	4	54	1	1	870	16	26	5	20	197	36	7	19	3	281
😢	0	0	7	18	1	186	3	0	0	390	0	7	7	8	108	84	1	0	1	28
🌹	2	3	3	5	6	17	223	2	0	1021	15	28	1	76	176	14	5	8	1	95
😊	0	0	8	13	7	0	39	73	0	637	5	13	2	50	55	62	5	5	3	211
😄	0	1	3	21	3	2	21	0	124	649	4	18	7	21	34	158	1	8	1	40
👍	1	3	13	43	12	8	104	0	0	8742	92	278	20	134	355	254	25	69	11	242
👑	0	0	3	4	1	0	21	0	0	587	399	31	3	17	38	21	0	8	1	41
👏	3	1	9	11	1	3	23	2	0	2515	41	320	8	26	89	103	3	19	1	93
👎	0	0	12	19	0	5	0	0	0	623	1	16	86	17	14	171	0	0	0	14
😊	0	2	7	45	6	3	107	3	1	1685	6	42	7	144	163	141	8	9	2	187
🙏	1	0	5	17	14	32	68	2	1	1438	11	34	12	64	1541	38	11	15	6	222
😂	2	0	25	49	0	12	3	0	2	1440	2	48	21	18	25	1111	4	5	2	45
✌️	1	0	2	5	5	0	4	1	0	426	3	7	4	2	16	28	59	1	1	392
💙	0	0	0	0	11	1	11	0	0	441	6	12	0	8	130	15	0	242	2	145
😍	1	0	2	7	4	2	31	1	0	844	24	44	2	29	64	79	4	16	22	137
❤️	5	0	6	16	19	10	55	1	1	1140	31	43	5	51	262	63	27	18	8	3889

Table 8: Error analysis of five emojis' comments with the percentage of misclassified comments that seem to fit the comments content in 30 randomly selected examples

#	Emoji	% in 30	Example #1	Example #2	Example #3
1	👑	100	את מלכה!!! at mlkh!!! you are a queen	מלדדדדדדדדך אוהבים אותך mlkkkkkkkkkk awhbim awtk king, we like you	יצאת נסיך icta nsik you came out a prince
2	🌹	96	מזל טוב mzl Twb congratulations!	שנה טובה šnh Twbh happy new year!	בוקר טוב וחג שמח bwqr Twb wxg šmx good morning and happy holiday
3	😢	83	יהי זכרם ברוך ihi zkrm brwk of blessed memory	עצוב מאוד `cwb mawd very sad	משתתפת בצערך mšttpt bc`rk to express one's condolence
4	🙏	73	החלמה מלאה hxlmh mlah full recovery	רפואה שלמה ובשורות טובות rpwah šlmh wbšwrwt Twbwt get well soon! may we hear only good news	יהי זכרם ברוך ihi zkrm brwk of blessed memory
5	👏	70	כל הכבוד מגיע לך kl hkbwd mgi` lk well done! you deserve it	אתה אלוף העולם ath alwp h`lm you are the world champion	בראבו brabw bravo

character n-grams representations are better than all the semantic vector representations, presented in Table 4.

The unigram representation outperforms all the other n-grams representations significantly. However, the best representations are the character 3-grams and character 4-grams. The F1 advantage of the character 3-grams representation over the unigram representation is statistically significant at the 0.01 level. This finding is interesting since the Hebrew root, which is the most basic form

of the word, to which other parts, such as affixes, can be added, contains three letters.

We also compared our results to a common FastText baseline [24], which is often on par with deep learning classifiers in terms of accuracy. Hyperparameters were set as default (as in Barbieri et al. [4]). We obtained an accuracy of 36.28, and precision, recall and a F1 scores of 0.2113 0.1504 0.1489, respectively. We observed that the character 3-grams representation outperforms the FastText baseline and its F1 advantage is statistically significant at the 0.01 level.

We note that in the SemEval shared task on multilingual emoji prediction the best performing system [14] on both English and Spanish datasets was based on text representations.

In Table 6, we show the precision, recall, and F1 measure of the character 3-grams representation for each emoji, sorted by the emoji frequency. Except for the 👏 emoji, the frequent emojis have higher F1 scores. Additional emojies with good performances are: 👑, 💙, and 😢. In general, the precision is higher than the recall. The classifier learned only part of the emojis features. Its default choice was 👍

4.3 Error Analysis

To better understand the challenges of the emoji prediction task, we analyzed the classification errors of the character 3-grams representation. First, in Table 7, we present the classification confusion matrix. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class.

Most of the classification errors were due to incorrect classification of comments into the most frequent emoji 👍. On average 55% of the wrong classification per emoji were classified as 👍. In the case of the emoji 👏 77% of the comments that contain it were classified as 👍. According to the Emojipedia [6] 👍 indicates approval and 👏 can be used as a round of applause, when used multiple times. Therefore, they might express a similar meaning.

12% of the comments with 😢 were wrongly classified as 🙏. It might mean being sad and praying for a quick recovery from illness. The 9% of the comments with 😢 which were wrongly classified as 😂 are probably a misinterpretation of the emoji as crying out of sadness instead of joy.

14% of the comments with 💙 and 16% of the comments with ❤️ were wrongly classified as ❤️. Due to the high frequency of the emoji ❤️, classifications error in the opposite direction were not frequent.

Next, in Table 8, we present five emojis that we have analyzed their misclassified comments. Since the emoji prediction for specific emojis could be subjective, we observed that in many of the cases both the classifier decision and the emoji that was chosen by the commentator seem to fit the comment content (the percentage of such cases in 30 randomly selected examples are given in the table). This can be explained by the finding of recent studies that people interpret emoji characters inconsistently [3, 33].

We note that in examples #3 and #5 there is a common expression יהי זכרם ברוך -ihi zkrm brwk (of blessed memory). This expression can express sadness (😢) or a prayer for the dead (🙏). In this paper, we simplified the prediction setting to single-label classification. However, in real life, it is a multi-label prediction task.

[6]https://emojipedia.org

5 CONCLUSIONS AND FUTURE WORK

We explored the task of emoji prediction as a single-label classification problem. We investigated four dimensional reduction methods which learn the distributional representations unsupervisingly using large unlabeled data. By applying 8 ML methods for emoji prediction, we compared the effectiveness of different semantic vector representations. The Word Embedding common representation for emoji prediction achieved the lowest F1 score. The ML algorithm that should be chosen depends on what you want to optimize; F1 or accuracy.

We showed that both of the text representations, i.e., n-grams and character n-grams, outperform all the semantic vector representations significantly. The best representation character 3-grams achieve an accuracy of 38.47% and a F1 score of 0.2176. It also significantly outperforms the efficient FastText algorithm.

In practice, more than one emoji may appear in a given text message. Thus, we plan to address the multi-label setting of the emoji prediction task. In addition, we plan to investigate the deep learning approach for our classification task as it has been shown to be effective [3, 12, 17, 42, 46]. Word Embedding is often the selected representation for neural network models. However, since it did not perform well in our setting, we would probably need to apply more sophisticated models than the deep learning baselines.

REFERENCES

[1] Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora Varvarigou. 2012. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 187–196.

[2] Francesco Barbieri, Luis Espinosa Anke, Jose Camacho-Collados, Steven Schockaert, and Horacio Saggion. 2018. Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4766–4771.

[3] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are Emojis Predictable?. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2. 105–111.

[4] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. 24–33.

[5] Francesco Barbieri, Luis Espinosa-Anke, and Horacio Saggion. 2016. Revealing patterns of Twitter emoji usage in Barcelona and Madrid. *Frontiers in Artificial Intelligence and Applications. 2016;(Artificial Intelligence Research and Development) 288: 239-44.* (2016).

[6] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 531–535.

[7] Francesco Barbieri, Luis Marujo, Pradeep Karuturi, William Brendel, and Horacio Saggion. 2018. Exploring Emoji Usage and Prediction Through a Temporal Variation Lens. *arXiv preprint arXiv:1805.00731* (2018).

[8] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis.. In *LREC*.

[9] Ella Bingham and Heikki Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 245–250.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[11] Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th evaluation campaign of natural language processing and speech tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018),*

Viviana Patti Tommaso Caselli, Nicole Novielli and Paolo Rosso (Eds.), Vol. 2263. 1–6.
[12] Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018. Contrastive Learning of Emoji-based Representations for Resource-Poor Languages. *arXiv preprint arXiv:1804.01855* (2018).
[13] Nurendra Choudhary, Rajat Singh, Vijjini Anvesh Rao, and Manish Shrivastava. 2018. Twitter corpus of Resource-Scarce Languages for Sentiment Analysis and Multilingual Emoji Prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1570–1577.
[14] Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 Task 2: SVMs perform better than RNNs in Emoji Prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. 34–38.
[15] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391–407.
[16] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. 48–54.
[17] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1615–1625.
[18] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint conference on*, Vol. 4. IEEE, 2047–2052.
[19] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl 1 (2004), 5228–5235.
[20] Gaël Guibon, Magalie Ochs, and Patrice Bellot. 2018. Emoji recommendation in private instant messages. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, 1821–1823.
[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[22] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 919–928.
[23] William B Johnson and Joram Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics* 26, 189-206 (1984), 1.
[24] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2. 427–431.
[25] Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25, 2-3 (1998), 259–284.
[26] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.
[27] Chaya Liebeskind, Shmuel Liebeskind, and Yaakov HaCohen-Kerner. 2017. Comment relevance classification in Facebook. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*. Budapest, Hungary.
[28] Chaya Liebeskind and Karine Nahon. 2018. Challenges in Applying Machine Learning Methods: Studying Political Interactions on Social Networks. In *Semantic Keyword-Based Search on Structured Data Sources*, Julian SzymaÅ„ski and Yannis Velegrakis (Eds.). Springer International Publishing, Cham, 136–141.
[29] Chaya Liebeskind, Karine Nahon, Yaakov HaCohen-Kerner, and Yotam Manor. 2017. Comparing Sentiment Analysis Models to Classify Attitudes of Political Comments on Facebook (November 2016). *Polibits* 55 (2017), 17–23.
[30] Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1520–1530.
[31] Chenglong Ma, Weiqun Xu, Peijia Li, and Yonghong Yan. 2015. Distributional representations of words for short text classification. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 33–38.
[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[33] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Eleventh International AAAI Conference on Web and Social Media*.
[34] Thomas Minka and John Lafferty. 2002. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 352–359.
[35] Dror Mughaz, Tzeviya Fuchs, and Dan Bouhnik. 2018. Automatic Opinion Extraction from Short Hebrew Texts using Machine Learning Techniques. *Computación y Sistemas* 22, 4 (2018).
[36] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one* 10, 12 (2015), e0144296.
[37] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
[38] Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond Just Text: Semantic Emoji Similarity Modeling to Support Expressive Communication. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 6.
[39] Stephan Raaijmakers and Wessel Kraaij. 2008. A shallow approach to subjectivity classification. In *ICWSM*.
[40] Francesco Ronzano, Francesco Barbieri, Endang Wahyu Pamungkas, Viviana Patti, Francesca Chiusaroli, et al. 2018. Overview of the EVALITA 2018 Italian Emoji Prediction (ITAMoji) Task. In *6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018*, Vol. 2263. CEUR-WS, 1–9.
[41] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.
[42] Toshiki Tomihira, Atsushi Otsuka, Akihiro Yamashita, and Tetsuji Satoh. 2018. What Does Your Tweet Emotion Mean?: Neural Emoji Prediction for Sentiment Analysis. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*. ACM, 289–296.
[43] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2016. Emojinet: Building a machine readable sense inventory for emoji. In *International Conference on Social Informatics*. Springer, 527–541.
[44] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. A semantics-based measure of emoji similarity. *arXiv preprint arXiv:1707.04653* (2017).
[45] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1, 6 (1945), 80–83.
[46] Chuhan Wu, Fangzhao Wu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Tweet Emoji Prediction Using Hierarchical Model with Attention. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 1337–1344.
[47] Ruobing Xie, Zhiyuan Liu, Rui Yan, and Maosong Sun. 2016. Neural Emoji Recommendation in Dialogue Systems. *arXiv preprint arXiv:1612.04609* (2016).
[48] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.
[49] Luda Zhao and Connie Zeng. [n. d.]. Using Neural Networks to Predict Emoji Usage from Twitter Data.

Emojis in Sentence Processing: An Electrophysiological Approach

Benjamin Weissman
bpweiss2@illinois.edu
University of Illinois at Urbana-Champaign, Department of Linguistics
Urbana, Illinois

ABSTRACT

This paper discusses two multi-experiment studies using the ERP methodology to investigate neural correlates of processing linguistic emojis. The first study examined the use of wink emojis used to mark irony and found the same ERP response complex that has been found in response to word-generated irony. Contingent upon individual differences in interpretation, these emojis are processed the same way as ironic words. The second study investigated the prediction of non-face emojis substituted for nouns. When predictability was high, unexpected emojis elicited the same ERP response patterns as words. Overall, the results of these two studies suggest that emojis used linguistically are processed in the same way as words and that individuals can integrate input from multiple modalities into a holistic representation of a single utterance.

KEYWORDS

emojis, ERP, irony, language processing, linguistics, multimodality, prediction, psycholinguistics

ACM Reference Format:
Benjamin Weissman. 2019. Emojis in Sentence Processing: An Electrophysiological Approach. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316544

The purpose of this paper is to introduce to the emoji research community a project on the psycholinguistic processing of emojis. As emojis can perform a range of linguistic functions, it is a worthwhile endeavor to investigate how users and receivers interpret these in real time. Especially with various public claims about emojis becoming 'their own language,' it is important to compare the processing of emojis to the processing of their word-based counterparts. A wide range of psycholinguistic methodologies could be utilized to investigate the processing of emojis in different sentence contexts; selection of an appropriate methodology is dependent on the aims of the research question and the specific emoji usage to be studied. Psycholinguistic investigations of emojis can uncover lots about emojis and about human language, including the speed and ease of emoji interpretation, the effects of context, regions of neural activation, acquisition of emoji competency, and even manifestations of aphasia.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316544

This paper presents two studies of the processing of emojis using the Event-Related Potentials (ERP) methodology. In this methodology, electrical activity from the brain is recorded via EEG while participants process auditory or visual stimuli. Neural activity is time-locked to the onset of meaningful stimuli and averaged across many trials and multiple participants to produce a picture of what happens, on average, in response to stimuli in different conditions. This averaged activity from electrodes of interest within a given time window is compared across conditions. ERPs are great for their millisecond-level temporal resolution and are therefore an ideal methodology to investigate the timing of cortical activity. ERP components are typically identified by their polarity (positive or negative), time course (milliseconds after critical stimulus onset), and scalp topography (where on the scalp the electrical activity is measured). An ERP effect is said to be significant when there is a significant difference between the average waveforms in two or more conditions during a given time window. The two studies presented here will provide information about when and how emojis in different language contexts are processed in the brain.

Study 1 is an investigation of the processing of ironic wink emojis and Study 2 examines the predictive processing of non-face emojis. Both of these studies are 'emoji versions' of experiments with findings that are established in the non-emoji linguistic literature, so the results of these studies are interpreted with respect to that previous work. As this paper is a synopsis of work and an entranceway to a discussion about studying the psycholinguistic processing of emojis, the results are not analyzed in full detail here. For a more detailed discussion of the results of Study 1, see the full write-up in Weissman & Tanner (2018) [4] or see Tanner, Goldshtein & Weissman (2018) [2] for a special look at individual differences in the study. The full-length paper for Study 2 is currently under review as Weissman, Cohn & Tanner (under review) [3].

Study 1 was a three-experiment project designed to investigate neural responses to wink emojis used as markers of irony. The consensus from previous ERP studies of word-generated irony is a P200 effect followed by a P600 effect in cases of irony. The P600, often discussed in linguistics, is a relatively late component that is typically conceived of as an index of reanalysis or reprocessing. The P200, much less discussed with respect to language processing, is an early component typically elicited when attentional resources are allocated. Materials for all three experiments were sentences with an emoji at the end. Sentences could be followed by an emoji that matched the valence of the sentence (positive or negative), an emoji that mismatched the valence of the sentence, or a wink emoji that rendered the sentence ironic. Examples are listed in Table 1 (emojis from EmojiOne [1]).

Table 1: Study 1 - Example Stimuli

Sentence	Condition
The stress was overwhelming 😕	Match
The stress was overwhelming 🙂	Mismatch
The stress was overwhelming 😉	Irony

Across the three experiments, both the P200 and P600 effects were reliably found to the ironic wink emojis. Since there was such a range of non-literal response rates to the comprehension questions, the relationships between ERP difference waves (formed by subtracting the Irony condition from the Match condition) and response rates across all three experiments were analyzed, uncovering a significant correlation with P600 amplitude ($r = .28, p = .004$). The higher an individual's non-literal response rate, the larger the P600 effect; this finding is consistent with the P600 as an index of reanalysis and indicates the robustness of the effect.

Study 2 was a two-experiment study focusing on cross-modal prediction of non-face emojis. Materials for this study consisted of sentences in which the last word could have an emoji substitution. Experiment 2a presented the sentences 'traditionally' with words only; Experiment 2b presented the sentences with the emoji substitutions. This study utilized two levels of sentence constraint: high and low, as determined by Cloze norming - a measure of predictability of a given word in a sentence. High constraint (high predictability) sentences could have either the expected completion or an unexpected yet plausible completion; low constraint sentences could have either a plausible (yet low predictability) or an implausible completion. Example stimuli are shown in Table 2.

Experiment 2a served as a (successful) replication of the previous studies that have examined the effects of sentence constraint, expectancy, and plausibility on traditional (word) sentences, finding an N400 effect to unexpected completions and an additional Late Frontal Positivity (LFP; 600-1000ms) when the completion is plausible albeit unexpected. The functional significance of the N400 effect is debated, but it is generally understood as index of the difficulty of semantic processing; the LFP has been suggested to be a reflection of discourse revision. The relevant effects in high constraint - the N400 and LFP - were significant to sentence-final emojis in Experiment 2b. Scalp distributions to the high constraint emoji completions, like those to word completions, were as expected for each effect - posterior for the N400, anterior for the LFP. In low constraint sentences, there was a significant difference between plausible and implausible emojis in the N400 time window, but this emoji-elicited negativity displayed a more anterior distribution than the word-based N400. It is hypothesized that high predictability, established during a sentence, facilitates the lexical access of emojis, hence the greater degree of similarity to word-elicited effects under high constraint.

A trial-level analysis found a significant correlation between the Cloze expectancy of a given completion and N400 amplitude to that emoji, replicating a correlation that has been shown to be present to words. Similar analyses were run with visual, conceptual, and semantic similarity (between expected emoji and actual emoji); that none of these correlations were significant indicates participants were anticipating lexical forms as opposed to visual, conceptual, or semantic features.

Table 2: Study 2 - Example Stimuli

Sentence	Constraint	Condition
She put more paper into the 🖨️	High	Expected
She put more paper into the 👞	High	Unexpected
His favorite animal is the 🐬	Low	Plausible
His favorite animal is the 🚒	Low	Implausible

A central finding across both studies is that emojis not only can serve linguistic purposes but also get processed in largely the same way as words when they do. This evidence counters various pop culture claims that emojis are 'their own language' and instead suggests a domain-general language processing system that can pull from multiple modalities. Wink emojis used ironically generated the same neural response as words used ironically, and emoji substitutions are lexically predicted and subject to the same constraint-based considerations as words. In both of these instances, emojis and sentences are combining to form a single, holistic utterance that pulls from two modalities; the neural evidence suggests that our mental representations of these occasions are holistic as well. Evidence from Study 2 suggests that high constraint facilitates the lexical access of emojis, indicating an interaction between the emojis and their linguistic environments.

These two studies have begun to provide insights into the neural correlates of processing emojis. With so many different uses of emojis and so much variability in their uses and interpretations, there is a lot of room for further research in the processing domain. The preponderance of ERP language studies allows for the comparison of many types of emoji uses to their non-emoji linguistic counterparts, including syntactic, lexical, and pragmatic phenomena. Less high-tech methodologies like eye-tracking and self-paced reading can also be used to investigate these topics, and more high-tech methods like fMRI can be used to provide more spatial information about neural activity in processing. Psycholinguistic methods can be used to investigate cross-cultural and cross-linguistic differences in emoji interpretations, including in second language learners. Studies of the processing of linguistic emojis in aphasic patients can reveal more about the language processing system in patients and in the nonaphasic population. Such explorations will not only advance our knowledge about emojis but also knowledge about our language processing and domain-general communicative systems.

ACKNOWLEDGMENTS

I would like to thank Darren Tanner, my PhD co-advisor and lab director. NSF grant BCS-1431324 awarded to DT funded the data collection of both of these projects. I would also like to thank multimodality expert Neil Cohn, a collaborator on Study 2.

REFERENCES

[1] 2019. EmojiOne. https://www.emojione.com/. (2019).
[2] Darren Tanner, Maria Goldshtein, and Benjamin Weissman. 2018. Individual differences in the real-time neural dynamics of language comprehension. In *Psychology of learning and motivation*. Vol. 68. Academic Press Cambridge, MA, 299–335.
[3] Benjamin Weissman, Neil Cohn, and Darren Tanner. under review. Lexical prediction of emojis in sentence contexts. (under review).
[4] Benjamin Weissman and Darren Tanner. 2018. A strong wink between verbal and emoji-based irony: How the brain processes ironic emojis during language comprehension. *PloS one* 13, 8 (2018), e0201727.

FATES on the WEB Chairs' Welcome

It is our great pleasure to welcome you to FATES on the Web 2019 - the *First Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web,* in conjunction with the Web Conference 2019!

The motivation for FATES is to bring together researchers and enthusiasts concerned with the urgent challenges concerning algorithmic fairness and accountability, transparency, and ethics on data management and social interaction on the Web, towards a Web that is truly inclusive and open.

We received 29 submissions covering a broad range of topics. We evaluated them regarding relevance, quality, and novelty. Each paper went through a double-blind review process and has received at least 3 reviews. We selected 4 full papers, 11 short papers, and 3 discussion papers, these last not included in the proceedings.

Submission Type	Reviewed	Accepted	
Full Technical Papers (up to 10 pages)	7	4	57%
Short Poster Presentations (up to 6 pages)	15	12	80%
Demos, Tutorials, or Others (up to 2 pages)	7	3	42%

We will have one keynote by Prof. Frauke Kreuter (Facebook and Maryland University) entitled "The Social Science of Privacy – Effects on Industry and Government Data Use"

Long papers will have 20 minutes of presentation, short and discussion papers 10 minutes.

We will have 7 technical sessions: two sessions group papers on Privacy, two on Bias, two on Social Media and one on Fairness.

We would like to sincerely thank The Web Conference organization for the constant help and support, the Program Committee members and reviewers for their excellent work and invaluable support during the review process, and the authors of the submitted papers for their very interesting and high-quality contributions. We also thank the ACM team for the proceedings. Most importantly, we thank all the attendees who ensure 'FATES on the Web' could be an appropriate forum of knowledge and experiences sharing about the new challenges in Social Web.

Hence we hope that you will find the workshop program interesting, providing you with a valuable opportunity to learn and share ideas with other researchers and practitioners from institutions around the world.

Chiara Renso
FATES Chair
ISTI/CNR, Italy

Daniel Sadoc Menasché
FATES Chair
Federal University of Rio de Janeiro, Brazil

Jonice Oliveira
FATES Chair
Federal University of Rio de Janeiro, Brazil

Lívia Ruback
FATES Chair
Federal University of Rio de Janeiro, Brazil

Carlos Castillo
FATES Chair
Universitat Pompeu Fabra, Spain

Jeanna Matthews
FATES Chair
Clarkson University, USA

FATES on the WEB Chairs' Welcome
Organization

Workshop/Track Chairs: Sihem Amer-Yahia (CNRS, Université Grenoble Alpes)
Mohammad Mahdian (Google)

Organization Committee: Chiara Renso, ISTI/CNR, Italy
Daniel Sadoc Menasché, Federal University of Rio de Janeiro, Brazil
Jonice Oliveira, Federal University of Rio de Janeiro, Brazil
Lívia Ruback, Federal University of Rio de Janeiro, Brazil
Carlos Castillo, Universitat Pompeu Fabra, Spain
Jeanna Matthews, Clarkson University, USA

Program Committee: *Aline Paes, UFF, Brazil*
Anastasios Giovanidis, LIP6/Sorbonne Univ. (CNRS), France
Ana-Andreea Stoica, Columbia University, USA
Anna Monreale, University of Pisa, Italy
Bettina Berendt, Univ. of Leuven, Belgium
Claudia Cappelli, PPGI/UFRJ, Brazil
Claudio Lucchese, Università Ca' Foscari Venezia, Italy
Cristina Muntean, ISTI/CNR, Italy
Daniel Schwabe, PUC-Rio, Brazil
Danilo Carvalho, Federal University of Rio de Janeiro
Elias Bareinboim, Purdue University, USA
Fabio Rangel, Federal University of Rio de Janeiro,
Fabrizio Silvestri, Facebook, London, UK
Giseli Rabello Lopes, Federal University of Rio de Janeiro, Brazil
Ida Mele, ISTI/CNR, Italy
Josep Domingo-Ferrer, Rovira i Virgili University, Catalonia
Karine Zeitouni, University of Versailles, France
Lily Hu, Harvard, USA
Marco Antonio Casanova, PUC-Rio, Brazil
Maria Luiza Machado Campos, Federal University of Rio de Janeiro, Brazil
Marília Guterres Ferreira, UDESC, Brazil
Michele Melchiori, University of Brescia, Italy
Raffaele Perego, ISTI/CNR, Italy
Regis Pires Magalhães, Federal University of Ceara, Brazil
Roberto Trani, ISTI/CNR, Italy
Vinicius Monteiro, UFPE, Brazil
Simone Diniz Junqueira Barbosa, PUC-Rio, Brazil
Stan Matwin, Dalhousie University, Canada

THE WEB
CONFERENCE

The Social Science of Privacy - Effects on Industry and Government Data Use

Frauke Kreuter
Univ. of Maryland, USA; Univ. of Mannheim; Institute for Employment Research, Germany

ABSTRACT

The sharing of data is at the core of many technology companies. Data sharing is also increasingly important for government decision-making, as stated by the Commission on Evidenced-Based Policymaking, which led to the Foundations for Evidence-Based Policymaking Act. However, in many instances, data used for decision-making is generated by people, and needs to be explicitly shared by the data subject with those wanting to use the data. The decision-making process behind sharing (private) information needs to be understood to assess (and circumvent) potential biases in the resulting data. When assessing bias in algorithmic decision-making, awareness of biases in the training data is essential. This presentation will review social science theories behind data sharing decision-making, highlight a series of experimental studies designed to affect sharing decisions, and present a framework design to detect sources of bias in various data sources.

KEYWORDS

data sharing, decision making, social science

ACM Reference Format:
Frauke Kreuter. 2019. The Social Science of Privacy - Effects on Industry and Government Data Use. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317600

SHORT BIO

Frauke Kreuter is a Professor in the Joint Program in Survey Methodology at the University of Maryland, Professor of Methods and Statistics at the University of Mannheim, and head of the statistical methods group at the German Institute for Employment Research in Nuremberg. Currently Frauke is a Visiting Scholar in Berkeley's Simons Institute Program on Privacy. Frauke serves on several advisory boards for National Statistical Institutes around the world. She is a Gertrude Cox Award winner, which recognizes statisticians in early- to mid-career who have made significant breakthroughs in statistical practice, winner of the inaugural Links Lecture Award and elected fellow of the American Statistical Association. Additionally, she is co-founder of the Coleridge Initiative, and founder of the International Program in Survey and Data Science.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317600

Algorithms for Fair Team Formation in Online Labour Marketplaces*

Giorgio Barnabò
Sapienza University of Rome, Italy
barnabo@diag.uniroma1.it

Adriano Fazzone
Sapienza University of Rome, Italy
fazzone@diag.uniroma1.it

Stefano Leonardi
Sapienza University of Rome, Italy
leonardi@diag.uniroma1.it

Chris Schwiegelshohn
Sapienza University of Rome, Italy
schwiegelshohn@diag.uniroma1.it

ABSTRACT

As freelancing work keeps on growing almost everywhere due to a sharp decrease in communication costs and to the widespread of Internet-based labour marketplaces (e.g., *guru.com*, *feelancer.com*, *mturk.com*, *upwork.com*), many researchers and practitioners have started exploring the benefits of outsourcing and crowdsourcing [13, 14, 16, 23, 25, 29]. Since employers often use these platforms to find a group of workers to complete a specific task, researchers have focused their efforts on the study of team formation and matching algorithms and on the design of effective incentive schemes [2–4, 17]. Nevertheless, just recently, several concerns have been raised on possibly unfair biases introduced through the algorithms used to carry out these selection and matching procedures. For this reason, researchers have started studying the fairness of algorithms related to these online marketplaces [8, 19], looking for intelligent ways to overcome the algorithmic bias that frequently arises. Broadly speaking, the aim is to guarantee that, for example, the process of hiring workers through the use of machine learning and algorithmic data analysis tools does not discriminate, even unintentionally, on grounds of nationality or gender.

In this short paper, we define the *Fair Team Formation problem* in the following way: given an online labour marketplace where each worker possesses one or more skills, and where all workers are divided into two or more not overlapping classes (for examples, men and women), we want to design an algorithm that is able to find a team with all the skills needed to complete a given task, and that has the same number of people from all classes.

We provide inapproximability results for the *Fair Team Formation problem* together with four algorithms for the problem itself. We also tested the effectiveness of our algorithmic solutions by performing experiments using real data from an online labor marketplace.

CCS CONCEPTS

• **Theory of computation** → *Design and analysis of algorithms.*

*The research for this work has been partially supported by the EU FET project MULTIPLEX 317532 and the ERC Advanced Grant 788893 AMDROMA "Algorithmic and Mechanism Design Research in Online Markets"

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317587

KEYWORDS

Fairness, Team Formation, Fair Set Cover, Outsourcing, Crowdsourcing

ACM Reference Format:
Giorgio Barnabò, Adriano Fazzone, Stefano Leonardi, and Chris Schwiegelshohn. 2019. Algorithms for Fair Team Formation in Online Labour Marketplaces. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317587

1 INTRODUCTION

An online labour marketplace is defined as a web application where workers can sell their services and skills in a fluid and delocalised fashion. Usually, employers pay workers hourly to complete a specific task without offering them any long-term employment arrangement. The OECD data on self-employment estimates that between 10% and 20% of workers in developed countries are self-employed, while it is estimated that in 2020, a full 40% of the US workforce will be freelancers [18].

While crowdsourcing adoption was driven, at least in part, by the assumption that problems can be decomposed into parts that can be addressed separately by independent workers, recent work suggests that crowdsourcing results can be improved by allowing some degree of collaboration among them [21, 26]. The idea of combining collaboration with crowdsourcing has led to research on *Team Formation* [1–3, 7, 9, 11, 15, 17, 20, 22, 28], in which a common thread is the need for complementary skills, and definitions differ in aspects such as objectives (e.g., load balancing and/or compatibility), constraints (e.g., worker capacity), and algorithmic set-up (online or offline).

As previously mentioned, these online marketplaces are largely managed through automatic algorithms designed to match supply and demand. Nevertheless, the objective of optimising a given task, which these algorithms are usually based on, goes openly against the need to ensure fairness and diversity, for example, in the composition of groups. We define *unfair discrimination* as treating someone differently on the base of his group membership, and not his merit. Since algorithms are "*black boxes*" usually protected by industrial secrecy, legal protections and even intentional obfuscation, most of the times discrimination becomes invisible, and mitigation impossible [12]. For this reason, data scientists and researchers have developed the *disparate impact* theory [8] whose aim is to spot unintended discrimination in algorithms outcomes. Among the many different sources of the bias on the Web, the one

that directly concerns us in the research of a solution for the *Fair Team Formation problem* is the *algorithmic bias*, that occurs when the bias is added by the algorithm itself or by the way this algorithm manages the bias present in the data it crunches.

Overview of problem setting and assumptions. In our framework, both workers and tasks are represented by sets of skills. Each skill of the task is possessed by at least one worker, while each worker has a defined cost and belongs alternately to one of two classes. In this setting, we consider the problem of finding the cheapest team of workers that together have all the necessary skills to complete the task, and that is made up of the same number of workers from both classes (fairness constrains). We call this general problem *Fair Team Formation*, which we formally define in Section 2 and solve in Sections 3-4.

Algorithmic techniques. To the best of our knowledge, we are the first to consider the *Fair Team Formation* problem, namely the weighed Set Cover problem with some fairness constraints imposed. As shown in section 3, the *Fair Team Formation* is NP-hard and inapproximable, for this reason the only thing we could do was to look for some algorithms that would function well in practical situations. Now, considering that our problem is closely related to the Set Cover problem [6, Chapter 35], it seemed natural to start from a reasoning similar to the one behind the Greedy Set Cover algorithm [27]. In the next sections, then, we will present four algorithms we developed to solve the *Fair Team Formation* problem: the first three are partially based on the Greedy Set Cover algorithm, while the fourth is a rounding algorithm based on the linear programming formulation of the Fair Team Formation problem.
Furthermore, since we are not able to calculate the value of the optimal solution in reasonable time, we have built a lower-bound for the cost of the optimal solution of the *Fair Team Formation* problem by solving the relaxed Linear Programming formulation of our problem. This lower-bound came in handy when we had to evaluate our algorithms performance.

Contributions. The key contributions of our work are:

- We formalise the *Fair Team Formation* problem, which is the problem of finding the cheapest team that can complete the task and, at the same time, that counts the same number of people from two not overlapping classes.
- We design four algorithms for solving our problem.
- We experiment on real data based on actual task requirements and worker skills from one of the largest online labor marketplaces, testing algorithms under a broad range of conditions.

2 PRELIMINARIES

In this section, we formally describe our setting and problem, and provide some necessary background.

2.1 Notation and Setting

Skills. We consider a set S of skills with $|S| = m$. Skills can be any kind of qualification a worker can have or a task may require, such as *video editing*, *technical writing*, or *project management*.

Tasks. We consider a set of $\mathcal{J}$ tasks (or jobs). Each task $J \in \mathcal{J}$ is independent, and requires a set of skills from S, therefore, $J \subseteq S$. In our setting we do not consider a streaming of tasks, but rather we take each task as a single instance of the problem.

Table 1: Notation

S	Set of skills, size m
$\mathcal{J}$	Set of tasks
$\mathcal{W}$	Set of workers, size n.
	$W^r_\ell = 1$ if worker r possess skill ℓ, 0 otherwise
P_ℓ	Subset of workers possessing skill ℓ
C	the two classes $\{class1, class2\}$

Workers. Throughout we assume that we have a set $\mathcal{W}$ of n workers: $\mathcal{W} = \{W^r;\ r = 1, \ldots, n\}$. Every worker r possesses a set of skills ($W^r \subseteq S$). Similarly to the tasks, we use W^r to denote both the worker and his/her skills. Moreover, each worker has a hiring cost, and belongs alternately to one of two not overlapping classes.

Classes. The workforce is split in two not overlapping classes $C = \{class1, class2\}$, for example women and men.

Coverage of tasks. Whenever task $J \subseteq S$ arrives, an algorithm has to assign one or more workers to it, i.e., a *team*. We say that J can be *completed* or *covered* by a team $Q \subseteq \mathcal{W}$ if for every skill required by J, there exists at least one worker in Q who possesses this skill: $J \subseteq \cup_{W \in Q} W$. We assume that for every skill in the incoming task there is at least one worker possessing that skill, so all tasks can be covered.

2.2 Problem Definition

We now define the problem that we study:

Problem 1 (The Fair Team Formation problem). *There exists a set of skills S. We have a pool of workers $\mathcal{W}$, where each worker $W^r \in \mathcal{W}$ is characterised by a subset of skills $W^r \subseteq S$, a hiring cost $c_r \in \mathbb{R}_{\geq 0}$, and belongs alternately to one of two not overlapping classes, $C = \{class1, class2\}$. Given a task $J \in \mathcal{J}$, the goal is to design an algorithm that, when task J arrives, decides which workers to hire such that all the tasks are covered by the workers who are hired, the total cost paid over all the tasks is minimised, and the team formed is made up of the same number of workers from both classes, $C = \{class1, class2\}$.*

One special case of the *Fair Team Formation problem*, where no fairness constraints are imposed, is the *Weighed Set Cover problem*. This problem can be effectively addessed through a greedy approach (see [30, Chapter 2]). As shown by Slavik [27], this greedy algorithm has an approximation ratio of $\log n - \log\log n + \Theta(1)$ [27]. Unfortunately, this result does not hold true for the Fair Set Cover that is the algorithmic core behind the Fair Team Formation problem.

3 INAPPROXIMABILITY & LOWER-BOUND

First, we will show that the Fair Team Formation problem (i.e. the Fair Weighted Set Cover problem) is inapproximable, then we will present two different lower-bounds that we can easily calculate, and use later on to evaluate the quality of the solutions found by our four algorithms.

3.1 Inapproximability of the Fair Set Cover Problem

Cover problems on hyper-graphs $H(V, E, w)$ aim to find a subset $S \subset E$ such that $v \in \cup_{S_i \in S} S_i$ for every $v \in V$ and w(S) is minimised. The vertex cover problem is a special case where we are given a graph $G(V', E')$ and aim to find a subset $S' \subset E'$ such that every edge $e \in E$ is incident to at least one node of S'. In terms of hyper-graphs, V corresponds to E' and each hyper-edge $h_v \in H$ corresponds to the set of edges incident to v.
Given a coloring $c : V \rightarrow \{red, blue\}$ of G, we consider a set of vertexes $S \subset V$ to be fair, if $|S \cap \text{RED}| = |S \cap \text{BLUE}|$. The fair vertex cover problem consists of finding a minimum vertex cover under the constraint that it is fair. Note that unlike the unconstrained fair vertex cover, such a set may not exist in general. Similarly, given a coloring of the sets $c : E \rightarrow \{red, blue\}$ the fair set cover problem consists of finding a minimum set cover $S \subset E$ such that $|S \cap \text{RED}| = |S \cap \text{BLUE}|$. We note that generally fair covers need not exist. This feature will allow us to show the following impossibility result.

Theorem 3.1. *Computing any finite approximation of the fair vertex cover problem is NP-hard.*

Proof. Let $G(V, E)$ be a graph, where we consider V to be red. Given an integer k, it is NP-hard to determine whether there exists a vertex cover of size at most k [10]. We add k blue vertexes V'. If there exists a fair vertex cover in $G'(V \cup V', E)$, then it can consist of at most k blue vertexes. Since any finite approximation of the fair vertex cover algorithm in particular determines the existence of a fair vertex cover, it also solves the decision problem of vertex cover. Hence, computing any finite approximation of the vertex cover is NP-hard. □

Corollary 3.2. *Computing any finite approximation of the fair set cover problem or the fair group Steiner tree problem is NP-hard.*

Proof. Both problems contain the vertex cover problem as a special case [10, 24]. □

Finally, it is worth noting that for the unweighted version of the Fair Set Cover problem (i.e. all workers have the same cost), and under the *enough workers* assumption, we can build a simple algorithm whose approximation factor is equal to $|C|H(|T|)$.

3.2 Lower-bound

When trying to solve an instance of the Fair Set Cover problem, we are often unable to calculate the value of the optimal solution in reasonable time; therefore, we are forced to use algorithms that find only a suboptimal solution to the problem. For this reason, it is important to have a lower-bound which we are sure that the value of the optimum would never go below. Obviously, a first really trivial lower-bound (TLB) is represented by the cost of the solution we obtain when the Greedy Set Cover is applied to the Fair Set Cover instance (after eliminating the fairness constraints), divided by its approximation factor; namely:

$$TLB = \frac{Cost(GreedySetCoverSolution)}{\log(n) - \log(\log(n)) + 3. + \log(\log(32)) - \log(32)} \quad (1)$$

A Lower-Bound from the Relaxed LP formulation of the Fair Set Cover problem. A computationally feasible and mathematically elegant way to calculate a better lower-bound for the Fair Set Cover problem is to solve its relaxed Linear Programming formulation. In a nutshell, we formulate the Fair Team Formation problem as an Integer Linear Programming problem, and then we relax its constraints. In this way, we obtain a Linear Programming problem that is solvable in polynomial time and whose solution always costs no more than the optimal solution that we would get if we were able to solve the integer linear programming. The Relaxed Linear Programming formulation of the Fair Set Cover problem is the following:

Relaxed Linear program for the Fair Set Cover problem:

$$\begin{cases} min & \sum_{i=1}^{|W|} c_i x_i & \\ s.t. & \sum_{i:s \in W_i} x_i \geq 1 & \forall s \in T \\ and & x_i \in [0,1] & \forall i \in \{1, ..., |W|\} \\ and & \sum_{i=1}^{|W|} k_i x_i = 0 & \end{cases}$$

Where x_i assumes either value 0 or 1, depending on whether the i_{th} worker is hired or not; c_i is the worker i hiring cost, and k_i is equal to -1 if worker belongs to *class*1, or to 1 if he belongs to *class*2.

4 THE FAIR TEAM FORMATION PROBLEM

Given the previous restrictive result, in this section, we provide four algorithms to solve the Fair Team Formation problem. Considering that the Fair Team Formation problem has a lot in common with the Set Cover problem, it seemed natural to start from a reasoning similar to the one behind the Greedy Set Cover algorithm. Therefore, algorithms 1, 2, 3 are partially based on the Greedy Set Cover algorithm, while algorithm 4 is a rounding algorithm based on the linear programming formulation of the Fair Team Formation problem. The only assumption we made is that there is always a team of workers that together have all the necessary skills to complete the task we are handling. In other words, the task is always coverable.

Fair Padding Greedy Set Cover algorithm. The first algorithm we came up with is a simple extension of the Greedy Set Cover algorithm where the cheapest workers of the class whose cardinality is lower are added to make the team fair. Algorithm 1 shows its pseudocode.

Algorithm 1 FairPaddingGreedySetCoverAlgorithm

Input: $(\mathbf{W}, J)$.
Ouput: FairTeam $W \subseteq \mathbf{W}$.

$W_0 \leftarrow$ GreedySetCoverAlgorithm(W, J)
if W_0 is not balanced (i.e. different number of workers from the two classes) **then**
  $W_1 \leftarrow W_0 \cup GetCheapestWorkers(W, W_0, MinorityClassCard.)$
  Return W_1
else
  Return W_0
end if

The time complexity of this algorithm is equal to the time complexity of the Greedy Set Cover algorithm, namely: $O(|\mathbf{W}||J|^2)$.

Fair Alternating Greedy Set Cover algorithm. Let's start by defining the marginal utility of each worker (WMU) as:

$$WMU = \frac{WorkerCost}{|SetOfWorkerSkills \cap SetOfTaskSkillsNotCoveredYet|} \quad (2)$$

Heuristically, at each stage, the AlternatingGreedySetCoverAlgorithm chooses the worker with the lower marginal utility alternating the class of workers within which it picks. Algorithm 2 shows its pseudocode.

Algorithm 2 FairAlternatingGreedySetCoverAlgorithm

Input: (W, J).
Ouput: FairTeam $W \subseteq \mathbf{W}$.

$W_1 \leftarrow$ AlternatingSetCover(W, J, StartingClass = 1)
if W_1 is not balanced (i.e. different number of workers from the two classes) **then**
  $W_1 \leftarrow W_1 \cup GetCheapestWorkers(W, W_1, MinorityClassCard.)$
end if
$W_2 \leftarrow$ AlternatingSetCover(W, T, StartingClass = 2)
if W_2 is not balanced (i.e. different number of workers from the two classes) **then**
  $W_2 \leftarrow W_2 \cup GetCheapestWorkers(W, W_2, MinorityClassCard.)$
end if
Return the cheapest team between W_1 and W_2

Also in this case, the time complexity is: $O(|\mathbf{W}||J|^2)$.

Fair Pairs Greedy Set Cover algorithm. Algorithm 3 is particularly simple and intuitive. Essentially, it is the application of the Greedy Set Cover algorithm to all possible pairs of workers. This idea has been suggested by [5]. Algorithm 3 shows its pseudocode.

Algorithm 3 FairPairsGreedySetCoverAlgorithm

Input: (W, J).
Ouput: FairTeam $W \subseteq \mathbf{W}$.

$W_{Pairs} \leftarrow$ PairsGenerator(W)
$W_0 \leftarrow$ CoupleGreedySetCover(W_{Pairs}, J)
Return W_0

Unlike the previous three algorithms, in this case the time complexity is: $O(|\mathbf{W}|^2|J|^2)$. The $|\mathbf{W}|^2$ factor is due to the fact that the greedy algorithm for the set cover problem has as input the set of all unordered couples of workers.

Relaxed Fair Set Cover Rounding algorithm. Algorithm 4 solves the relaxed linear programming formulation of the Fair Team Formation problem assigning to each worker a real number between 0 and 1: this number could be interpreted as the worker's probability to be hired. Then, it continues by creating random teams of workers using these probabilities until it finds a team that is both fair and able to complete the task.

Algorithm 4 RelaxedFairSetCoverRoundingAlgorithm

Input: $(\mathbf{W}, J)$.
Ouput: FairTeam $W \subseteq \mathbf{W}$.

$Hiring_{ProbabilityVector} \leftarrow$FairTeamFormationRelaxedLP($\mathbf{W}$, J)
$\mathbf{W}_{Sorted} \leftarrow SortAccordingToProbabilityVector(\mathbf{W})$
while ¬(W balanced ∧ task skills are all covered) **do**
  $W \leftarrow EmptyTeam$
  for $w \in \mathbf{W}_{Sorted}$ **do**
    add w to W with probability equal to $Hiring_{ProbabilityVector}(w)$
    if (W balanced ∧ task skills are all covered) **then**
      Return W
    end if
  end for
end while
Return W

5 EXPERIMENTS

In this section, we will present some experiments that we ran on a real dataset to evaluate the algorithms' performance by comparing their cost.

5.1 The Freelancer dataset

To create a large pool of tasks and workers needed to test the algorithms, we decided to use a dataset obtained from *Freelancer.com*: the largest online marketplace for outsourcing in its category according to data from Alexa (Feb. 2018). The input data that we obtained contain anonymised profiles from people registered as freelancer in this marketplace. This includes their self-declared sets of skills, as well as the average rate that they charge for their services. Data have been cleaned to remove skills that were not possessed by any worker, and skills that were never required by any task. Concerning tasks, we had access to a large sample of tasks commissioned by buyers in the marketplace. Some relevant characteristics of our data are summarised in table 2.

Table 2: Freelancer Dataset Characteristics

Dataset	Freelancer
Skills (m)	175
Workers (n)	1,211
Tasks (T)	992
...distinct	600
Average Skills/Worker	1.45
Average Skills/Task	2.86

As shown above, our dataset contains 992 tasks, but since many of them require exactly the same set of skills we decided to take into account only the 600 distinct tasks. The average number of skills per worker is 2.86 and the maximum is 6 skills.

Experiments Design. In the first place, we split the 1211 Freelancer workers into two different classes, and we considered six different compositions of the two groups. In brief, we used a random procedure to select respectively 10%, 30%, and 50% of all workers, and we assigned these workers to one of the two classes, while the remaining to the other. After that, for each of these configurations, we ran the four algorithms we designed to solve the Fair Team Formation Problem, obtaining fair teams to complete each of the 600 tasks.

5.2 Experiments

As shown in figure 1, we observe a shift to the left in the distribution, as the workforce becomes more balanced. In most cases the price of the fair team is no more than four times the value of the best lower bound (LB), although for a few tasks the FairAlternatingGreedySetCoverAlgorithm finds solutions that are even eight times the value of the lower bound. It is also worth noting that the progressive balancing of workers' colours has a significant effect on all algorithms, except for the RelaxedFairSetCoverRoundingAlgorithm whose cost (cost_RLP) distribution remains more or less consistent as the workforce changes. Moreover, from figure 1 we can see that all distributions are concentrated around a value of 2, indicating that our algorithms have an heuristic approximation ratio of 2, at least on this specific dataset.

In summary, histograms in figure 2 give us some important information about the overall algorithms performance, obtained by choosing the less expensive fair team among the four on a case-by-case basis. The balance between the two classes of workers does not influence the cost distributions suggesting that some algorithms are able to efficiently address the problem of strong unbalances between the two groups of workers; second, we can observe that the best solution cost is never more than four times the value of its best lower bound, and it rarely exceeds a factor of two.

To conclude, the RelaxedFairSetCoverRoundingAlgorithm beats them all: it was able to find a team whose cost is equal to the best solution cost in no less than 66% of cases, and with an average success rate of 85% (all configurations of colours considered). On the contrary, the FairPaddingGreedySetCoverAlgorithm always had the worst overall performance, never reaching a success rate higher than 70%.

6 CONCLUSIONS & FUTURE WORK

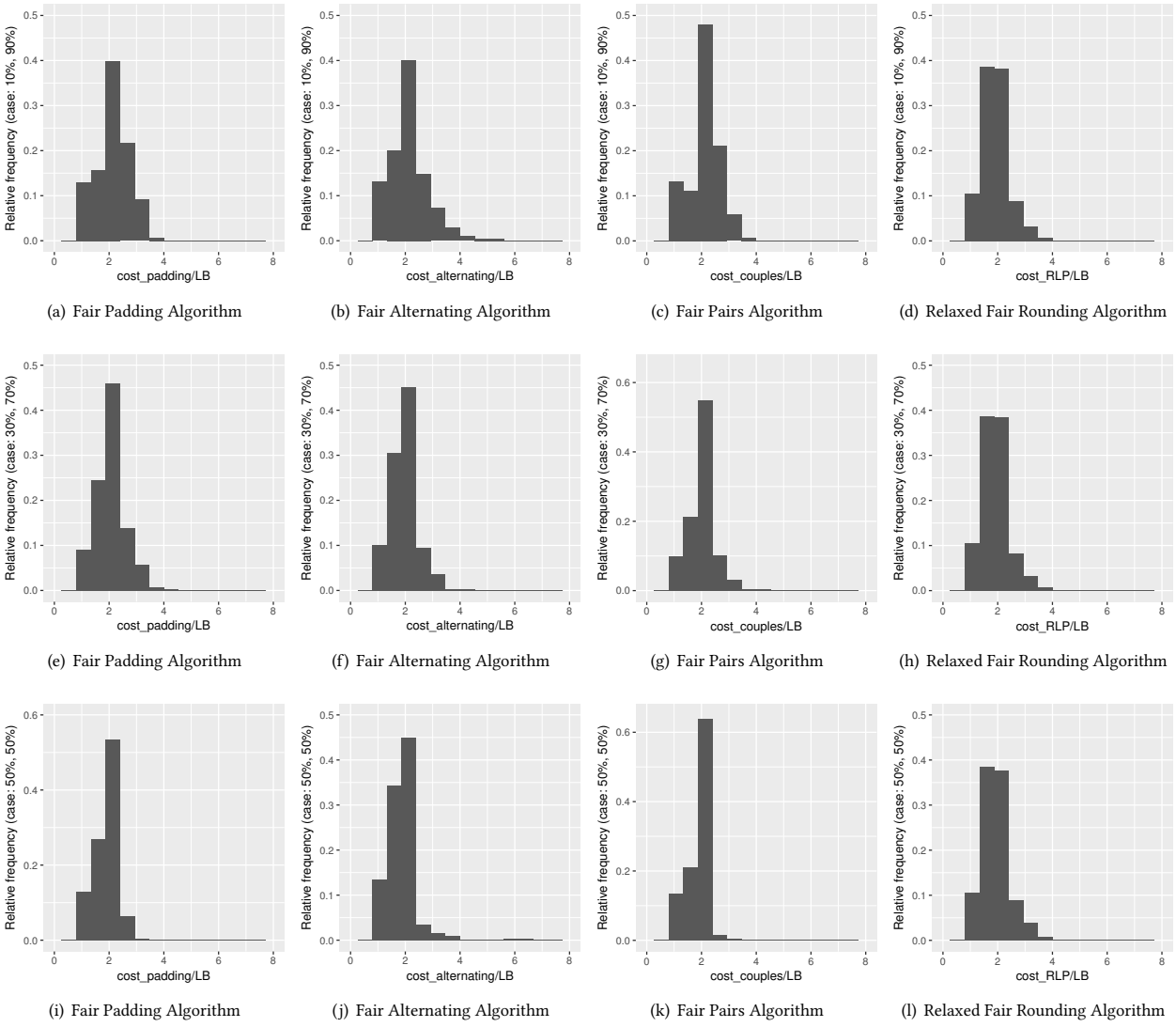
In this work, we have defined the Fair Team Formation problem, that is a variation of the Set Cover problem where each subset is assigned a colour, and whose goal is to find the cheapest collection of subsets that both covers the input set, and that is made up of the same number of subsets of each colour.

Despite the discovered inapproximability results, in particular for the Fair Team Formation problem, we have focused our research on the design and implementation of four algorithms for that problem, and we have also tested them on a real dataset. From the experiments we conducted on the Freelancer dataset, it turned on that the FairAlternatingGreedySetCoverAlgorithm and the RelaxedFairSetCoverRoundingAlgorithm outperform both the FairPaddingGreedySetCoverAlgorithm and the FairPairsGreedySetCoverAlgorithm in almost every case we considered, both in terms of solutions cost and in terms of solutions size. Overall, among these four algorithms, it seems more reasonable to opt for the RelaxedFairSetCoverRoundingAlgorithm. We can conclude that, even if the problem is not approximable in its weighted version, the algorithms we designed could be effectively used in practical contexts, and are able to find good solutions to many instances of the problem, at least in the limited case presented in the experiments chapter.

Throughout this paper, we assumed that all workers in the workforce can be hired to complete each task; in other words, when creating a team for any task, algorithms can pick team members among all workers who make up the workforce: this is a pretty strong and unrealistic assumption since usually workers have a limited available time; therefore, in the future, it could be interesting to extend this research further by considering a stream of tasks, or by limiting the number of times each worker can be hired. Now, coherently with the scientific literature on Team Formation, another possibility worth of some consideration is the introduction of a social network among workers; this would lead to the emergence of new interesting research questions, such as finding a fair team that minimises the distance among workers. Finally, to make this research more exhaustive, it could be convenient to study how the behaviour of Fair Team Formation algorithms changes with different datasets. Our experiments were, in fact, limited to tasks of no more than 6 skills, and to a workforce of only 1211 workers. We think that moving forward with this research could lead to some really useful and interesting results that, in turn, could help marketplaces designers, as well as policy makers, to better engineering, managing, and regulating these platforms.

REFERENCES

[1] A. An, M. Kargar, and M. ZiHayat. Finding affordable and collaborative teams from a network of experts. In *SDM*, pages 587–595, 2013.
[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Power in unity: forming teams in large-scale community systems. In *ACM CIKM*, pages 599–608, 2010.
[3] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *WWW*, pages 839–848, 2012.
[4] A. Anagnostopoulos, C. Castillo, A. Fazzone, S. Leonardi, and E. Terzi. Algorithms for hiring and outsourcing in the online labor market. In *KDD*, 2018.
[5] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. *CoRR*, abs/1802.05733, 2018.
[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009.
[7] C. Dorn and S. Dustdar. Composing near-optimal expert teams: A trade-off between skills and connectivity. In *OTM Conferences (1)*, pages 472–489, 2010.



(a) Fair Padding Algorithm (b) Fair Alternating Algorithm (c) Fair Pairs Algorithm (d) Relaxed Fair Rounding Algorithm

(e) Fair Padding Algorithm (f) Fair Alternating Algorithm (g) Fair Pairs Algorithm (h) Relaxed Fair Rounding Algorithm

(i) Fair Padding Algorithm (j) Fair Alternating Algorithm (k) Fair Pairs Algorithm (l) Relaxed Fair Rounding Algorithm

Figure 1: Distribution of solutions cost over best lower bound for three different class balances.

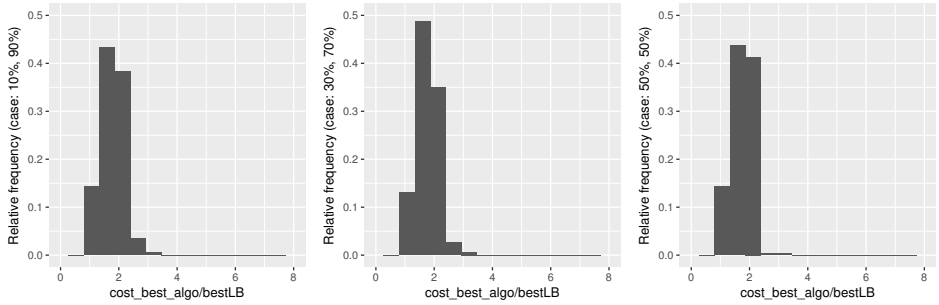


Figure 2: Distribution of best solution cost over best lower bound for three different class balances.

[8] M. Feldman, S. A. Friedler, J. Moeller, C. E. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.
[9] A. Gajewar and A. D. Sarma. Multi-skill collaborative teams based on densest subgraphs. In *SDM*, pages 165–176, 2012.
[10] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979. ISBN 0-7167-1044-7.
[11] B. Golshan, T. Lappas, and E. Terzi. Profit-maximizing cluster hires. In *ACM SIGKDD*, pages 1196–1205, 2014.
[12] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 2125–2126, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.
[13] J. Howe. The rise of crowdsourcing. *WIRED (June)*, 2006.
[14] L. Jeppesen and K. Lakhani. Marginality and problem-solving effectiveness in broadcast search. *Organization Science*, 21(5):1016–1033, 2010.
[15] M. Kargar and A. An. Discovering top-k teams of experts with/without a leader in social networks. In *CIKM*, pages 985–994, 2011.
[16] A. Kittur, B. Smus, S. Khamkar, and R. Kraut. Crowdforge: Crowdsourcing complex work. In *Annual ACM Symposium on User Interface Software and Technology*, pages 43–52, 2011.
[17] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *ACM SIGKDD*, pages 467–476, 2009.
[18] G. Laumeister. The next big thing in e-commerce: Online labor marketplaces, 2014. Last accessed on 2018-08-18.
[19] K. Levy and S. Barocas. Designing against discrimination in online markets. *Berkeley Tech. LJ*, 32:1183, 2017.
[20] C.-T. Li and M.-K. Shan. Team formation for generalized tasks in expertise social networks. In *SocialCom/PASSAT*, pages 9–16, 2010.
[21] A. Majchrzak and A. Malhotra. Towards an information systems perspective and research agenda on crowdsourcing for innovation. *The Journal of Strategic Information Systems*, 22(4):257–268, 2013.
[22] A. Majumder, S. Datta, and K. V. M. Naidu. Capacitated team formation problem on social networks. In *KDD*, pages 1005–1013, 2012.
[23] T. Malone, R. Laubacher, and C. Dellarocas. The collective intelligence genome. *MIT Sloan Management Review*, 51(3):21–31, 2010.
[24] G. Reich and P. Widmayer. Beyond steiner's problem: A VLSI oriented generalization. In *Graph-Theoretic Concepts in Computer Science, 15th International Workshop, WG '89, Castle Rolduc, The Netherlands, June 14-16, 1989, Proceedings*, pages 196–210, 1989.
[25] D. Retelny, S. Robaszkiewicz, A. To, W. Lasecki, and J. Patel. Expert crowdsourcing with flash teams. In *ACM symposium on User interface software and technology*, pages 75–85, 2014.
[26] C. Riedl and A. W. Woolley. Teams vs. Crowds: Incentives, member ability, and collective intelligence in temporary online team organizations. 2016.
[27] P. Slavık. A tight analysis of the greedy algorithm for set cover. *Journal of Algorithms*, 25(2):237–254, 1997.
[28] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *ACM SIGKDD*, pages 939–948, 2010.
[29] J. Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Anchor Books, 2004.
[30] V. V. Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2013.

Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Daniel Borkan
dborkan@google.com
Jigsaw

Lucas Dixon
ldixon@google.com
Jigsaw

Jeffrey Sorensen
sorenj@google.com
Jigsaw

Nithum Thain
nthain@google.com
Jigsaw

Lucy Vasserman
lucyvasserman@google.com
Jigsaw

ABSTRACT

Unintended bias in Machine Learning can manifest as systemic differences in performance for different demographic groups, potentially compounding existing challenges to fairness in society at large. In this paper, we introduce a suite of threshold-agnostic metrics that provide a nuanced view of this unintended bias, by considering the various ways that a classifier's score distribution can vary across designated groups. We also introduce a large new test set of online comments with crowd-sourced annotations for identity references. We use this to show how our metrics can be used to find new and potentially subtle unintended bias in existing public models.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Model verification and validation**.

ACM Reference Format:
Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317593

1 INTRODUCTION

Evaluating the predictions of models is an essential part of machine learning work, and selecting the appropriate evaluation metrics for a given task is a central question for researchers in this field. As machine learning models are being used for an ever-expanding set of use cases, many have raised concerns about the potential negative impact of unintended identity-based bias that could be present in the models [3, 8]. In recent years, significant research has been published presenting new metrics intended to measure this unintended bias during model evaluation [4, 7, 10].

Our interest is in improving text classification models used to identify toxicity in comments from online discussions, but the evaluation methods presented here can be applied to a broad range of classification applications. "Toxicity", defined as anything that is *rude, disrespectful, or unreasonable that would make someone want to leave a conversation*, is an inherently complex and subjective classification task. Machine learning systems, if not constrained, will often learn the simplest associations that can predict the labels, so any incorrect associations present in the training data can produce unintended associations in the final model. Toxicity models specifically have been shown to capture and reproduce biases common in society, for example mis-associating the names of frequently attacked identity groups (such as "gay", and "muslim" etc.) with toxicity [5, 17]. This unintended model bias could be due to the demographic composition of the online user pool, the latent or overt biases of those doing the labelling, or the very selection and sampling process used to choose which items to label.

Regardless of the source, we focus on measuring a particular aspect of model unfairness - the skewing of classifier scores, and thus output labels, due to identity related content within the text. We use a definition of model fairness similar to *equality of odds* defined in [10]. As in that work, we assume the existence of a test set with reliable labels across a range of groups. Given such a test set, we consider unintended bias to be present in the model if the model performance, according to relevant performance metrics, varies across the set of designated groups. It is important to highlight that the assumption of reliable labels is significant and doesn't hold in all use cases. We mitigate the impact of this assumption by demonstrating our results against both a synthetic test set with labels that are constructed to be reliable and a large human-annotated test set with high rating redundancy.

We propose a suite of threshold agnostic performance metrics to measure the extent of unintended model bias. Many prior methods for measuring unintended bias in classification systems rely on selecting a threshold, a choice that can drastically change results. In practice, classification models often return *scores* instead of binary classification decisions, enabling them to be used with a variety of thresholds or for ordering data. For these models, threshold dependant metrics can obscure the view of unintended bias and thus be misleading to practitioners. Threshold agnostic metrics capture the behavior of the underlying model itself, and thus can allow a more comprehensive comparison of the model's performance and limitations.

While the practice of building machine learning models is simplest when there are single metrics for model comparison, unintended bias in models can be extremely nuanced and varied across

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317593

groups and so single metrics are likely to obfuscate essential information. We therefore propose a suite of five metrics, derived from ROC-AUC, Equality Gap, and Mann-Whitney U metrics, each of which captures a different aspect of model performance, and a different potential type of unintended bias. Viewing a suite of metrics across a range of groups will provide much greater insight into the nuance of unintended bias, and hopefully reveal new opportunities for mitigating these biases.

We apply these metrics with two test sets, again making the assumption that the labels are reliable. One is a synthetic test set, identical to the one presented in [5]. The other, introduced in this work, is a new human-labeled dataset of nearly 2 million comments, specifically created for evaluation of unintended bias[1]. This includes 450,000 comments annotated with the identities that are referenced in the text.

We demonstrate our proposed metrics and datasets on two publicly accessible models that are trained to detect toxicity in text (provided by the Perspective API [11].) One of these models is claimed to be trained using a bias mitigation technique, as described in [5] and [13]. We show that our metrics and datasets illuminate unintended bias in the original model and provide new insight into the effectiveness of the bias mitigation between the models.

2 RELATED WORK

Significant recent work has been published on defining how the concepts of fairness and unintended bias apply to machine learning models. Much of this work provides metrics to quantify the presence of unintended bias according to specific definitions. In [8], [12], and [15], one can read overviews of several of these metrics and the relationship and trade-offs between these various metrics. An industry wide push towards increased transparency about machine learning data sources, techniques, and evaluation criteria, as advocated in [16], has underscored the importance of choosing the right metrics.

Recent works such as [4] and [3] demonstrate the value of these metrics to highlight the potential for unfair impact with the use of machine learning models in applications. In addition, several works, such as [10], [1], [6] and [5] provide techniques intending to mitigate bias in machine learning models. While out of the scope of this work, effective unintended bias measurement is essential to measure progress in bias mitigation.

[17] focuses on measuring and reducing the gender-based bias specifically. In addition to using an approach from [5], they introduce measure of the false positive and negative equality gaps, very similar to [1], that measure the different between the true positive rates between the subgroup and the overall background.

This is a relaxation of the *equalized odds* fairness constraint presented in [10] that can serve as a metric only for a classifier that produces binary labels. However, many models produce a probability distribution. As these probability scores may be used in a variety of ways or with a variety of thresholds, the Equality Gap falls short for many purposes, as it evaluates the model only at one specific threshold.

[1]https://git.io/fhpcC

The related work on counterfactual fairness [9] proposes a number of techniques to reduce classifiers affinity for assigning attributed features to identity terms. As before, the equality gap is measured for a particular threshold chosen to maximize the accuracy over a designated test set.

[5] introduces a threshold agnostic metric for unintended bias, but a follow up work by the same authors, [2], highlights several limitations of this metric. Specifically, the metric is not robust to variations in the class distribution between different identity groups. In addition, with a single metric, some important information may be hidden as different types of bias could obscure one another.

Our proposed metrics differ from these early approaches because they are threshold agnostic, robust to class imbalances in the dataset, and because they provide more nuanced insight into the types of bias present in the model, as we will see in Section 3.3.

3 NEW METRICS

3.1 AUC-Based Metrics

In this section we introduce three new metrics to measure unintended bias, based on the Area Under the Receiver Operating Characteristic Curve (ROC-AUC, or AUC) metric. For any classifier, AUC measures the probability that a randomly chosen negative example will receive a lower score than a randomly chosen positive example, i.e. that the two will be correctly ordered. An AUC of 1.0 means that all negative/positive pairs are correctly ordered, with all negative items receiving lower scores than all positive items.

A core benefit of AUC is that it is threshold agnostic. An AUC of 1.0 also means that it is possible to select a threshold that perfectly distinguishes between negative and positive examples, i.e. that the classes are perfectly separable via the model score.

Most metrics for unintended bias rely on dividing the test data up by identity or demographic based subgroups and computing metrics for each group. For our metrics, we also divide data by subgroup. However, instead of calculating metrics on the subgroup data exclusively, our metrics compare the subgroup to the rest of the data, which we call the "background" data.

Consider the hypothetical score distributions for an example model shown in Figure 1. The background score distributions are shown on the top, with negative examples in green on the left and positive examples in pink on the right. The second (bottom) distributions shows scores for examples within a specific identity subgroup. We can see clearly that the examples within the identity receive higher scores, both for positive and negative examples. This score shift is one way that unintended bias can manifest in a model. In fact, many types of unintended bias can be uncovered by looking at differences in the score distributions between background data and data from within a specific identity (although not all differences imply harmful bias). Using three new metrics based on AUC, we can specifically measure variations in the distributions that cause misorderings between negative and positive examples, i.e. variations that limit the possibility of selecting a single effective threshold.

In the example shown in Figure 1, the dataset is divided into background and identity subgroups and negative and positive classifications, creating four distinct subsets: negative examples in the background, positive examples in the background, negative examples in the subgroup, and positive examples in the subgroup.

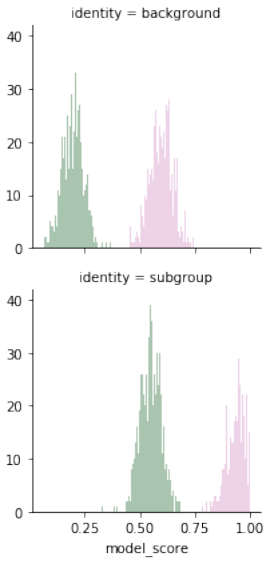


Figure 1: Hypothetical model score distributions for the background data (top) and one identity subgroup (bottom).

We define three AUCs to measure negative/positive mis-ordering between these four subsets.

Definition 1. *Let D^- be the negative examples in the background set, D^+ be the positive examples in the background set, D_g^- be the negative examples in the identity subgroup, and D_g^+ be the positive examples in the identity subgroup.*

$$\text{Subgroup AUC} = \text{AUC}\,(D_g^- + D_g^+) \quad (1)$$

$$\text{BPSN AUC} = \text{AUC}\,(D^+ + D_g^-) \quad (2)$$

$$\text{BNSP AUC} = \text{AUC}\,(D^- + D_g^+) \quad (3)$$

Table 1 illustrates how each of these three AUCs is constructed by looking at a different subset of the data. Each of these three AUCs captures a unique and specific aspect of the model performance:

Subgroup AUC Equation term 1, the calculates AUC on only the examples from the subgroup. This represents model understanding and separability within the subgroup itself.

Background Positive Subgroup Negative (BPSN) AUC Equation term 2 calculates AUC on the positive examples from the background and the negative examples from the subgroup. This value would be reduced when scores for negative examples in the subgroup are *higher* than scores for other positive examples, as in the example in Figure 1. These examples would likely appear as *false positives* within the subgroup at many thresholds.

Background Negative Subgroup Positive (BNSP) AUC Equation term 3 calculates AUC on the negative examples from the background and the positive examples from the subgroup. This value would be reduced when scores for positive examples in the subgroup are *lower* than scores for other negative examples. The examples would likely appear as *false negatives* within the subgroup at many thresholds.

Looking at these three metrics together for any identity subgroup will reveal how the model fails to correctly order examples in the test data, and whether these mis-orderings are likely to result in false positives or false negatives when a threshold is selected.

An important quality of the AUC metric is that it is robust to data imbalances in the amount of negative and positive examples in the test set. This is especially relevant when measuring unintended bias, because in real-world data, the amount of examples in each identity subgroup, and the balance between negative and positive examples can vary widely across groups (in fact, this variation is often a source of bias). Enforcing that for each AUC, either all negative or all positive examples (or both in Subgroup AUC) come from one identity group, means that mis-orderings involving that particular subset cannot be drowned out by results from other groups, ensuring that these metrics are robust to data imbalances likely to occur in real data.

3.2 Average Equality Gap

We now introduce two additional threshold agnostic metrics, building from a strict generalization of the Equality Gap metric.

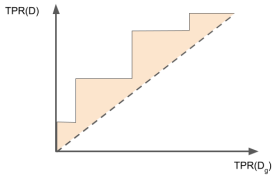


Figure 2: An illustrative plot of the true positive rate of the subgroup and background distributions of a hypothetical classifier. This shaded area can be captured by the AEG.

The Equality Gap is the difference between the true positive rates of the subgroup ($\text{TPR}(D_g)$) and the background ($\text{TPR}(D)$), at a specific threshold. Consider 2, which plots these rates against each other for every possible threshold t, for some hypothetical classification model. The hypothetical classifier is biased against the subgroup as $\text{TPR}(D_g) < \text{TPR}(D)$ at different levels at different thresholds. The shaded area captures the average bias across all thresholds for the classifier.

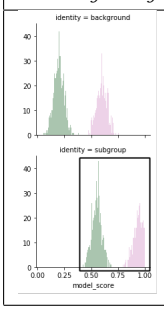
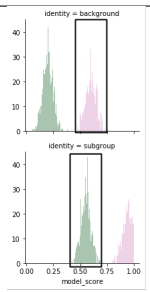
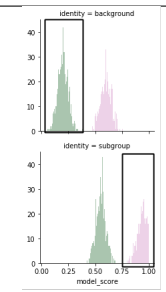
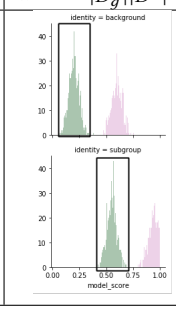
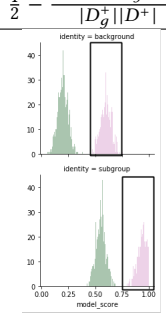
Definition 2 (Positive Average Equality Gap). *For each threshold, t, if you plot the true positive rate of the subgroup as $x(t)$ and the true positive rate of the background as $y(t)$ then the Positive Average Equality Gap is the area between the curve $(x(t), y(t))$ and the line $y = x$, i.e.*

$$\text{Positive AEG} = \int_0^1 (y(t) - x(t))\, dt \quad (4)$$

There is also the analogous definition with true negative rates in place of true positive ones. This would correspond to the other component of equalized odds. We call this the *Negative Average Equality Gap*. Together, we call the pair the Average Equality Gap metrics (AEGs).

3.2.1 Relationship to Mann-Whitney U Metrics. Another way to generalize the Equality Gap metric is from the perspective of the separability of the score distributions, similar to the AUC metrics in the previous section. With the AUC metrics, we measured

Table 1: The full suite of unintended bias metrics. Highlighted distributions indicate which data is used to calculate each metric.

Unintended Bias Metrics				
AUCs			Average Equality Gaps	
Subgroup AUC	Background Positive Subgroup Negative (BPSN) AUC	Background Negative Subgroup Positive (BNSP) AUC	Negative AEG	Positive AEG
$\text{AUC}(D_g^- + D_g^+)$	$\text{AUC}(D^+ + D_g^-)$	$\text{AUC}(D^- + D_g^+)$	$\frac{1}{2} - \frac{\text{MWU}(D_g^-, D^-)}{\lvert D_g^- \rvert \lvert D^- \rvert}$	$\frac{1}{2} - \frac{\text{MWU}(D_g^+, D^+)}{\lvert D_g^+ \rvert \lvert D^+ \rvert}$
				

mis-orderings between positive and negative examples across the subgroup and background, with the goal of few mis-orderings or high separability. For the AEGs, we compare positive examples from the subgroup with positive examples from the background, with the goal of *low* separability. In particular, if Equality of Opportunity held for our classifier at every threshold then that would mean that, if a point i (with model score $\hat{Y}_i$) were chosen uniformly at random from the positive background data D^+ and a point j (with model score $\hat{Y}_j$) were chosen uniformly at random from our subgroup data D_g^+, then:

$$P\left\{\hat{Y}_i > \hat{Y}_j | Y_i \in D^+, Y_j \in D_g^+\right\} = \frac{1}{2}$$

That is to say, given that both data points are positive examples from the two distributions, the probability that either score is higher than the other should be the same. Again, as a metric, we express this as:

Definition 3 (Positive Average Equality Gap). *If a point i (with model score $\hat{Y}_i$) were chosen uniformly at random from the background data D^+ and a point j (with model score $\hat{Y}_j$) were chosen uniformly at random from our subgroup data D_g^+, then the average equality gap is:*

$$\text{Positive AEG} = \frac{1}{2} - P\left\{\hat{Y}_i > \hat{Y}_j | Y_i \in D^+, Y_j \in D_g^+\right\}$$

We can rewrite the above definitions using the Mann-Whitney U test statistic [14] to come up with an efficient closed form for computing the Positive Average Equality Gap:

Definition 4 (Positive Average Equality Gap).

$$\text{Positive AEG} = \frac{1}{2} - \frac{\text{MWU}(D_g^+, D^+)}{|D_g^+||D^+|}$$

Analogously, the equivalent definitions of Negative Average Equality Gap follows from simply substituting the negative datasets for the positive ones in the above definitions.

Finally, we show:

Theorem 3.1. *All definitions of Average Equality Gap are equivalent.*

Proof. The equivalence of definitions 3 and 4 follow from the definition of the Mann-Whitney U test statistic. The equivalence of 4 and 2 is shown identically to earlier proofs of the equivalence of AUC and MWU scores (see, for example Section 2 of [14]). □

3.2.2 Properties of Average Equality Gap. By Definition 2, we see that the Average Equality Gap as a metric can range in value from -0.5 to 0.5. At each of these extremes, it represents a different type of bias where the TPR of the subgroup is consistently higher or lower, respectively, than that of the background.

The optimal value of the Average Equality Gap metric is 0, which means the subgroup and background distributions have identical means.

Definition 2 leads to the following corollaries:

Theorem 3.2. *If Equality of Opportunity holds for every threshold then the Average Equality Gap will be 0.*

Proof. If Equality of Opportunity holds for every threshold then $x(t) = y(t)$ for all t in Definition 2. Thus the Average Equality Gap will be 0. □

Theorem 3.3. *If the Average Equality Gap is 0 then Equality of Opportunity must hold for some non-trivial threshold $0 < t < 1$.*

Proof. Equality of Opportunity always trivially holds if you threshold the classifier at 0 or 1. However, if the Average Equality Gap is 0 then, by the Average Value Theorem for Integrals, we must also have that $x(t) = y(t)$ for some $0 < t < 1$ where $x(t), y(t)$ are defined as in Definition 2. □

3.3 Comparison of Metrics

In this section, we discuss the different strengths and weaknesses of the AUC-based metrics and the AEGs at tackling a range of common biases. Table 2 outlines simulated data distributions that exhibit common biases and reports our newly introduced metrics for bias.

- **A** **Small score shift** This type of bias occurs when a machine learning classifier outputs a consistently higher (or lower) score for a subgroup than for the overall data distribution. We refer to this as a "small" score shift as the shift is not to the extent as to confuse negative examples from the subgroup with positive examples from the background distribution (or vice versa). Thus, it is still possible to choose a single threshold that achieves a perfect separation of positive and negative examples for both the subgroup and the background. As we can see from Table 2, the Average Equality Gap (AEG) is the only metrics which pick up this subtle form of bias.
- **B** **Large score shift** This type of bias is similar to the previous one, but the shift is now large enough that negative examples from the subgroup are mis-ordered with positive examples from the overall distribution. Thus, it is no longer possible for a single threshold to separate positive and negative examples from the subgroup or the background. A threshold that is ideal for the background distribution would result in false positives within the subgroup, captured by the low value in BPSN AUC. This score shift is also detected by both AEG metrics, which are higher here than in **A** capturing the greater extent of the bias in this scenario.
- **C** **Score shift and size skew (more positive)** The bias shift in this distribution is the same as the previous one, but the available subgroup data is skewed toward having many more positive than negative examples. Again, this picked up by the BPSN and BNSP AUCs and AEGs, exactly as in **B**, demonstrating the metrics' robustness to imbalanced datasets.
- **D** **Left score shift** This type of bias is similar to **B**, except that the subgroup scores are shifted downward (to the left). Here, the bias is captured again by negative values in both AEGs, indicating the downward shift in scores, and by the low value in BNSP AUC, indicating the likelihood of false negatives for the subgroup at thresholds ideal for the background.
- **E** **Low subgroup separability** This type of bias represents a classifier which simply underperforms on a subgroup relative to the background distribution, resulting in low separability within the subgroup only. The intermingling of positive and negative examples from the subgroup is captured in the Subgroup AUC metric. Because this also implies a shift in one or both of the subgroup distributions, it will also be captured by the AEG metrics. The sign of the AEG metric corresponds to the direction of the shift, so in this example, the -0.48 corresponds to the left-shift of the positive examples in the subgroup and the 0.48 corresponds to the right shift of the negative examples in the subgroup. Note that the shifts are not so large as to cause overlap between negative examples in the subgroup and positive examples in the background (or vice versa), so the BPSN and BNSP AUCs are both still high.
- **F** **Wide subgroup score range without overlap** This distribution represents the scenario where the classifier produces a higher variance of scores for the subgroup than the background distribution, but the means for these distributions are the same. None of the bias metrics introduced here perceive any bias in this case. Whether this is considered problematic bias depends on the use case, if it is, you'll want to use more sensitive metrics than those introduced in this work.
- **G** **Wide subgroup score range with overlap** This distribution is similar to **F**, except that the subgroup distributions are now so wide that they overlap with each other and with the opposite class background distributions. The AUC metrics, especially Subgroup AUC, perceive this type of bias.

In summary, the different metrics can be used in combination to diagnose different types of bias. Of course, these scenarios are not an exhaustive list of all possible types of bias, but they help to illuminate some of the differences between these metrics.

Overall, Subgroup AUC and BPSN and BNSP AUCs identify any bias significant enough to cause mis-orderings between negative and postive examples, i.e. bias that interferes with selecting a single threshold that works similarly across groups. Subgroup AUC highlights when those mis-orderings are caused by poor model understanding within the subgroup, and BPSN and BNSP AUCs highlight when the misorderings are caused by score shifts. The AEGs go beyond the AUCs to identify bias in the distribution itself, even when (non-trivial) perfect thresholding is possible. Both AEGs and BPSN and BNSP AUCs provide insight into the directionality of score shifts.

It's important to note that the correct handling of subtle variations in distributions must be decided on case by case basis. Some cases of subtle variations that the metrics highlight, such as **A**, may not be considered problematic bias in all model use cases. Other, even more subtle variations like **F**, do not trigger any of these metrics. Generally, if the AUC values reveal variation across groups, that likely indicates a problematic bias as it confirms mis-orderings around the relevant subgroups. If the AEGs reveal variation across groups, that *may* indicate problematic bias, however that bias may not be severe enough to cause mis-orderings. And, as with any other suite of metrics, it's always possible that there is subtle bias that the metrics cannot detect.

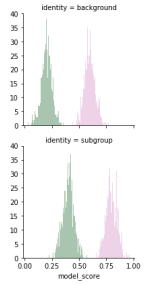
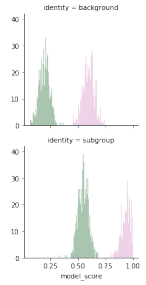
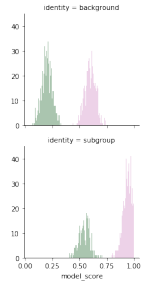
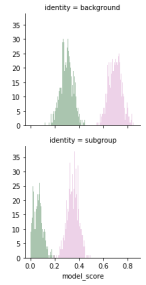
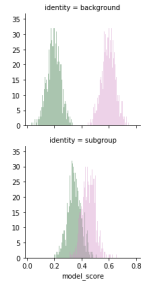
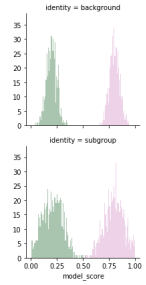
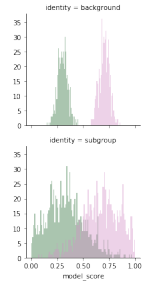
4 EXPERIMENTAL RESULTS

We demonstrate this suite of metrics using the publicly available toxicity classifiers provided by the Perspective API ([11]). We use two test sets, 1) a synthetically generated, bias-focused test set following [5] and 2) a large dataset of online comments with human labels for both identity and toxicity.

4.1 Models

Using our metrics, we compare two versions of Perspective API's toxicity classifier, the initial TOXICITY@1 and the latest TOXICITY@6. TOXICITY@1 was shown to have significant unintended bias around identity words like "gay" and "transgender", both by independent analysis and by the Perspective team [13]. TOXICITY@6 was built using the bias mitigation techniques presented

Table 2: Examples of score distributions demonstrating bias.

	A	B	C	D	E	F	G
Description	Small right score shift	Large right score shift	Score shift and size skew (more positive)	Large left score shift	Low subgroup separability	Wide subgroup score range without overlap	Wide subgroup score range with overlap
Subgroup AUC	1.0	1.0	1.0	1.0	0.93	1.0	0.92
BPSN/BNSP AUCs	0.99/1.0	0.76/1.0	0.76/1.0	1.0/0.77	0.99/0.99	1.0/1.0	0.98/0.97
AEGs	0.42/0.42	0.5/0.5	0.5/0.5	-0.5/-0.5	-0.48/0.48	-0.02/0.02	0.03/-0.03
							

in [5] and, and therefore we expect to see reduced unintended bias between these two models across our new metrics.

4.2 Synthetic Test Set

The synthetic dataset contains 77k examples generated from templates using 50 identity terms, 50% toxic and 50% non-toxic across all terms. These examples are constructed explicitly to measure unintended bias based on identity terms. The examples are simple sentences that should be clearly toxic or clearly non-toxic, regardless of identity terms present.

4.3 Synthetic Test Set Results

In Table 3, we show Subgroup AUC, BPSN AUC, BNSP AUC, Negative AEG, and Positive AEG for both TOXICITY models on the synthetic dataset. The dataset contains 50 identity terms, here we show results for the lowest performing 20 subgroups.

Bias tends to skew towards toxicity for certain groups. For TOXICITY@1, we see low values for BPSN AUC for identity terms *homosexual*, *gay*, and *lesbian*, and to a lesser extent *transgender* and *heterosexual*. This reveals the tendency of the TOXICITY@1 model to mis-associate those words with toxicity, and therefore produce high toxicity scores for non-toxic examples with these words, aligning with the findings of frequent false positives for these identity terms in [13]. This unintended bias is reduced in TOXICITY@6, but not completely eliminated.

Subgroup AUC and BNSP AUC show relatively high values across all groups, in both TOXICITY@1 and TOXICITY@6. This emphasizes that the model is generally effective at distinguishing toxic from non-toxic examples within every group (subgroup AUC), even for the groups that show an incorrect tendency towards toxicity in the BPSN AUC discussed above. In addition, the high BNSP values show that mis-orderings of toxic comments referring to identities that would lead to false negatives are rare.

Score distributions vary widely across groups. The Average Equality Gap values reveal a much more complex view of the model than the AUCs. As expected, identities with low BPSN AUC (e.g. *homosexual*, *gay*, and *lesbian*) also have high Negative AEG values, both indicating an upward shift (towards toxicity) of model scores for non-toxic items. However, several identities, such as *bisexual*, *trans*, *queer*, and *black*, have high BPSN AUCs and also high Negative AEGs values. For these groups, there is also an upward shift in model scores for non-toxic items, but it is not large enough to cause mis-orderings with toxic items. In addition, we see relatively large improvement in BPSN AUC values from TOXICITY@1 to TOXICITY@6, but only minimal change in the Negative AEGs for the same identities, confirming again that the upward shift of model scores for non-toxic comments with these identities is reduced but not eliminated.

4.4 Human Labeled Dataset

Synthetic test sets, while useful for capturing issues not present in real data, may not provide accurate results for real scenarios with different data distributions. In addition, synthetic sets are limited to the specific identity terms that are manually curated, and therefore are unlikely to capture the true diversity of ways that identities are discussed in real conversation.

To facilitate unintended bias evaluation on real data, we designed techniques to have humans label the identity content within real data. We presented crowd raters with a comment and asked a set of questions including, for example, “What genders are referenced in the comment?” and “What races or ethnicities are referenced in the comment?”. For each question, raters selected the set of identities present in the comment from a provided list. Using human labeling for identity content allows us to capture nuanced identity content that term-based analysis would miss, such as “same-sex marriage” or “people who do not believe in any god”.

The set of identities labelled by raters is not comprehensive and does not provide universal coverage. This set was designed to balance the coverage of identities, crowd rater accuracy, and ensure that each labeled identity has enough examples in the final data set

Table 3: Comparison between TOXICITY@1 and TOXICITY@6 on the synthetic dataset.

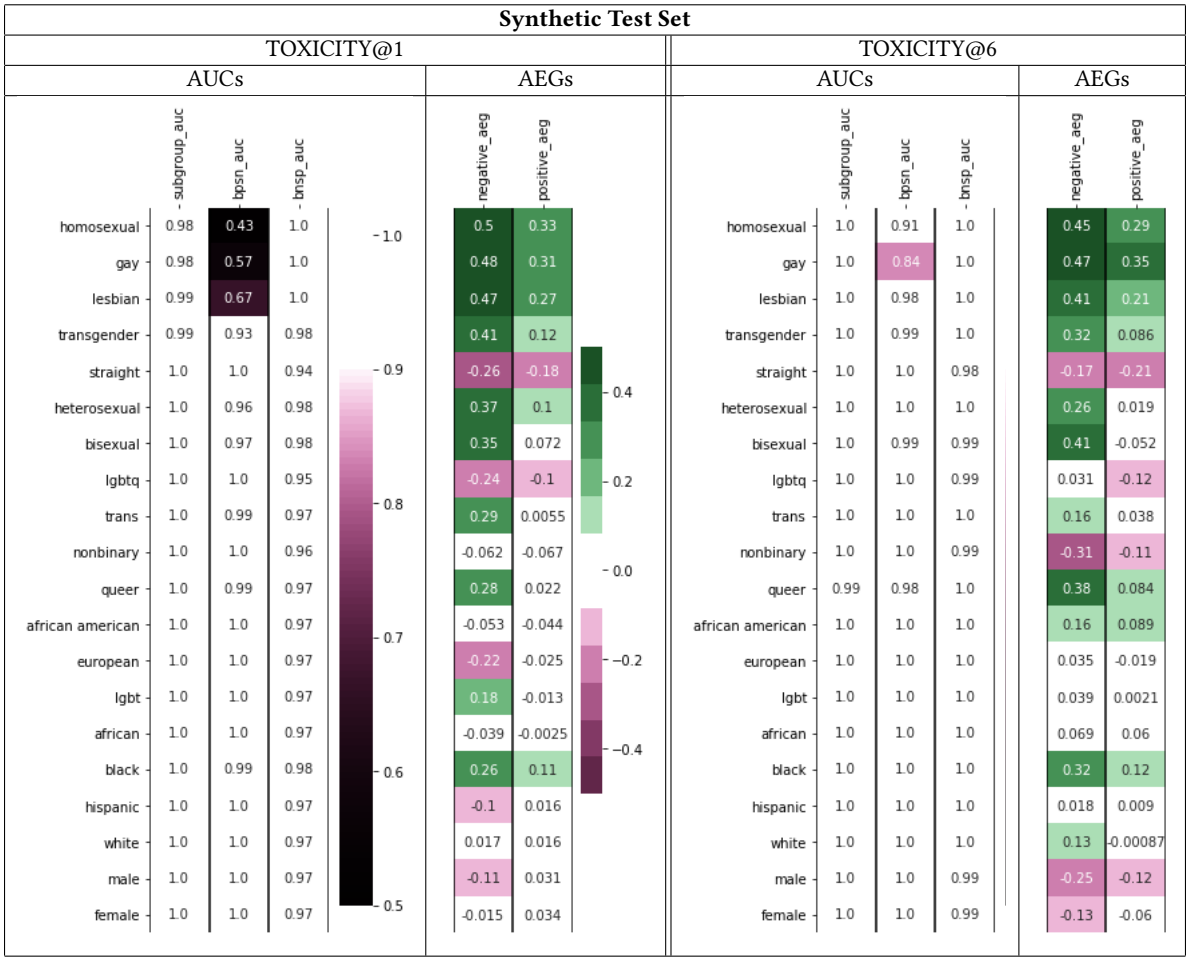


Table 4: Full list of identity options presented to crowd raters

Category	Identity Options
Gender	Male, Female, Transgender, Other gender
Sexual Orientation	Heterosexual, Homosexual, Bisexual, Other sexual orientation
Religion	Christian, Jewish, Muslim, Hindu, Buddhist, Atheist, Other religion
Race or ethnicity	Black, White, Latino/Latina/Latinx, Other race or ethnicity
Disability	Physical disability, Intellectual or learning disability, Psychiatric disability or mental illness, Other disability

to provide meaningful results. The full list of identities labeled can be found in Figure 4.

This data was also labeled for toxicity using the same crowd rating guidelines as published by the Perspective API ([18], [19]). This labeling asks raters to rate the toxicity of a comment, selecting from "Very Toxic", "Toxic", "Hard to Say", and "Not Toxic". Raters were also asked about several subtypes of toxicity, although these labels were not used for the analysis in this work.

Using these rating techniques we created a dataset of 1.8 million comments, sourced from online comment forums, containing labels for toxicity and identity. While all of the comments were labeled for toxicity, and a subset of 450,000 comments was labeled for identity. Some comments labeled for identity were preselected using models built from previous iterations of identity labeling to ensure that crowd raters would see identity content frequently.

Table 5 shows the toxicity percentage for a selection of identities, illustrating that there is an imbalance in toxicity between different identities, emphasizing the value of metrics that are robust to these data skews.

To enable further research in this field, this entire dataset and annotations will be released under a Creative Commons license at https://git.io/fhpcC.

4.5 Human Labeled Dataset Results

Applying the AUC and AEG metrics to this real dataset reveals several new insights about the two toxicity models. Table 6 compares

Table 5: Percentage of toxic comments by identity in the human labeled dataset for a selection of identities.

Subgroup	Count	Percent Toxic
all comments	1,804,875	8.00%
male	44,484	15.03%
female	53,429	13.68%
transgender	2,499	21.29%
heterosexual	1,291	22.77%
homosexual	10,997	28.38%

results for both TOXICITY@1 and TOXICITY@6 on all metrics, for both short comments (less than 100 characters) and all comments.

We present results on short comments separately because, according to [13], the bias mitigation implemented between TOXICITY@1 and TOXICITY@6 focused on short comments. And, because, as we will see, results vary between short and long comments. The identities shown in Table 6 are all identities that contained more than 100 examples of short comments. Walking through Table 6, we highlight several important findings.

Real data reveals more unintended bias than synthetic data. Comparing the real data results to the synthetic data results in Table 3, we find lower values and more variation across identity subgroups in the real data than we do in synthetic data. However, the identity terms that show the lowest model performance in the synthetic data are aligned with the identity groups that show the lowest performance in the real data. The synthetic data is intentionally very simple, so it is best at revealing very large discrepancies in performance that are tied very narrowly to specific identity terms, while the real data is much more broad and nuanced, but also potentially noisier.

Bias tends to skew towards toxicity. Across both models and both short and long comments, we see lower values for Subgroup AUC and BPSN AUC and higher values for BNSP AUC. We also tend to see positive values for Negative AEG and negative values for Positive AEG. Together, all of these metrics indicate that the models have a tendency to skew non-toxic comments that discuss identity towards toxicity. Given that the domain is online discussions, the tendency towards toxicity for identity-related discussion is aligned with general societal perceptions of online conversation, and therefore may not be surprising. In fact, we even see that the identities which have the most bias towards toxicity are identities that are frequently attacked in online discussion, such as *black*, *homosexual_gay_or_lesbian*, and *jewish* and the identities which have the least bias towards toxicity are *male*, *christian*, and *asian*, again aligned with societal stereotypes. From these results and from what we know of society, we can conclude that for modeling toxicity in online conversation, the risk for bias skewed towards toxicity for some groups is high. However, it's still important to include all metrics, as bias can appear in unexpected places.

Short comments show expected bias mitigation. Looking at the AUCs and AEGs for short comments, TOXICITY@6 outperforms TOXICITY@1 in most identities and metrics. This is expected since TOXICITY@6 includes data based on the bias mitigation methods described in [5] and [13]. That work found that imbalances in toxicity in the training data for certain identity words were a major source of bias, and that those imbalances were more prevalent among short comments. To mitigate the bias, the authors added additional training data to even out the prevalence of toxicity for those identity words, adding the most additional training data to short comments, where the imbalance was largest, so it is unsurprising that the mitigation would have more impact on short comments. Finally, the identity groups which do not improve in these metrics, such as *black* and *asian*, were not associated with any additional non-toxic training data.

For all comments, we see much less change between TOXICITY@1 and TOXICITY@6, which raises the potential that the bias mitigation did not generalize to longer comments. However, this dataset also reveals that there was *more* unintended bias among short comments in TOXICITY@1, exactly as predicted by the training data analysis in [5]. This is evident by the fact that that for short comments the AUC values are lower (especially Subgroup AUC and BPSN AUC) and all metrics have more variation across groups.

Table 7 visualizes the larger bias for short comments and the improvement between TOXICITY@1 and TOXICITY@6 for the identity group *homosexual_gay_or_lesbian*. Each image shows the score distributions for non-toxic comments (green arc, towards the left) and toxic comments (purple arc, towards the right). The top charts show the score distributions for both models on short comments and the lower charts shows all comments. For short comments and TOXICITY@1, the non-toxic and toxic distributions for the identity group are almost entirely overlapping with each other, and with the background toxic distribution, illustrating the bias we see in the low Subgroup AUC, BPSN AUC, and Negative AEG. For short comments and TOXICITY@6, the non-toxic distribution for the identity group is shifted to the left (lower scores) showing some improvement, which aligns with the improvement we see in those metrics.

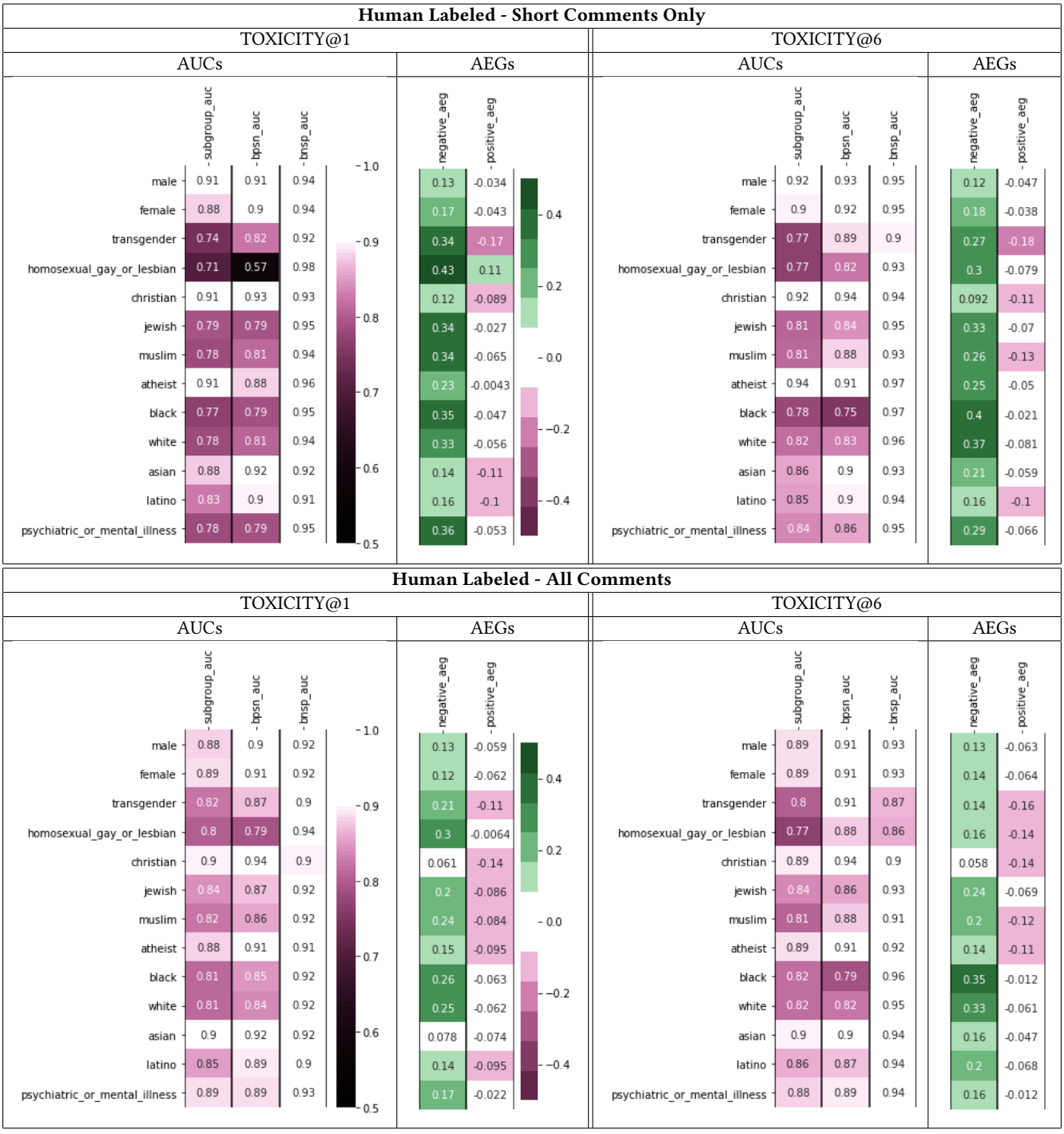
The metrics and distributions show that the bias mitigation brought unintended bias in short comments, where it was strongest, in line with unintended bias overall, where it was less visible. These results mirror the bias-mitigating training data adjustment, where toxicity percentages among short comments with certain identity terms were brought in line with percentages for other words. Overall, this evaluation on real data reveals that the bias mitigation between TOXICITY@1 and TOXICITY@6 impacts the model as designed, but there is still room for improvement.

5 CONCLUSION AND FURTHER WORK

We introduced a new suite of metrics for unintended bias, based on ROC-AUC and Mann-Whitney U scores. These metrics provide a detailed and nuanced view of the types of bias present in a model and overcome limitations of similar metrics like Equality Gap in that they are threshold agnostic.

We developed and applied an evaluation method for our introduced metrics using a variety of example illustrative distributions. This highlights the differences in various metric behaviors for different kinds of bias. We then demonstrated our metrics using existing toxicity classifiers that are provided by the Perspective API [11].

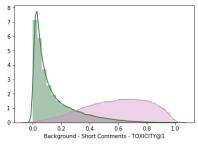
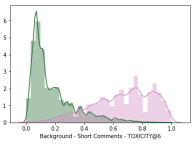
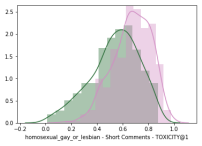
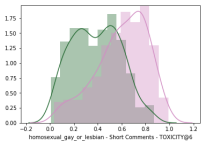
Table 6: Comparison between TOXICITY@1 and TOXICITY@6 short comments and all comments in the human labeled dataset.

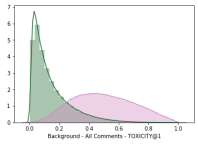
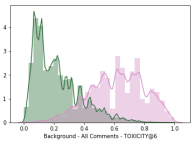
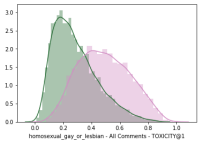
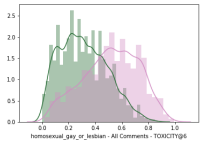


This involved adapting existing synthetic datasets used for unintended bias measurement of text classifiers.

Finally we extend beyond the synthetic test set methodology, leveraging the improved nuance of the newly introduced metrics by crowdsourcing a large new corpus of nearly 2 million annotations of comments, providing one of the first studies of unintended bias based on identity references in text classification on real data. Our evaluation using this new dataset highlights how the new metrics

Table 7: Score distributions comparison for comments labeled as containing identity *homosexual* for short comments only and all comments, comparing TOXICITY@1 and TOXICITY@6.

Short Comments Only	
TOXICITY@1	**TOXICITY@6**
Background	**Background**
	
homosexual_gay_or_lesbian	**homosexual_gay_or_lesbian**
	

All Comments	
TOXICITY@1	**TOXICITY@6**
Background	**Background**
	
homosexual_gay_or_lesbian	**homosexual_gay_or_lesbian**
	

also reveal new challenges for bias mitigation, highlighting that bias is still present in models that have undergone some bias mitigation.

Further work in this area could study:

- Developing effective strategies for choosing optimal thresholds to minimize unintended bias. While the threshold agnostic metrics we present provide an understanding of bias in the underlying model scores, this does not mean that all thresholds will have the same results.
- Evaluating the relative benefit of the newly introduced dataset compared to sub-string matching of terms that reference an identity.
- A more systematic definition of the kinds of synthetic distributions that can be used to evaluate and categorize metrics for unintended bias.
- Developing a full taxonomy of different possible biases and a systematic approach for these metrics to be used in their diagnosis.

REFERENCES

[1] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075 (2017). http://arxiv.org/abs/1707.00075

[2] Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Limitations of Pinned AUC for Measuring Unintended Bias. *CoRR* (2019).

[3] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior* 36, 1 (2009), 21–40. https://doi.org/10.1177/0093854808326545 arXiv:https://doi.org/10.1177/0093854808326545

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[5] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

[6] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11–21. http://aclweb.org/anthology/D18-1002

[7] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[8] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016). http://arxiv.org/abs/1609.07236

[9] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2018. Counterfactual Fairness in Text Classification through Robustness. *CoRR* abs/1809.10610 (2018). arXiv:1809.10610 http://arxiv.org/abs/1809.10610

[10] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *CoRR* abs/1610.02413 (2016). http://arxiv.org/abs/1610.02413

[11] Jigsaw. 2017. Perspective API. https://www.perspectiveapi.com/

[12] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR* abs/1609.05807 (2016). http://arxiv.org/abs/1609.05807

[13] Lucy Vasserman, John Li, CJ Adams, Lucas Dixon. 2018. Unintended bias and names of frequently targeted groups. https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23

[14] S. J. Mason and N. E. Graham. [n. d.]. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128, 584 ([n. d.]), 2145–2166. https://doi.org/10.1256/003590002320603584 arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/003590002320603584

[15] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 107–118. http://proceedings.mlr.press/v81/menon18a.html

[16] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[17] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2799–2804. http://aclweb.org/anthology/D18-1302

[18] Nithum Thain, Lucas Dixon, and Ellery Wulczyn. 2017. Wikipedia Talk Labels: Toxicity. (2 2017). https://doi.org/10.6084/m9.figshare.4563973.v2

[19] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1391–1399.

Collaborative Explanation of Deep Models with Limited Interaction for Trade Secret and Privacy Preservation

Josep Domingo-Ferrer
Universitat Rovira i Virgili
Dept. of Computer Science and Mathematics
UNESCO Chair in Data Privacy
CYBERCAT-Center for Cybersecurity Research of Catalonia
Tarragona, Catalonia
josep.domingo@urv.cat

Cristina Pérez-Solà
Universitat Rovira i Virgili
Dept. of Computer Science and Mathematics
UNESCO Chair in Data Privacy
CYBERCAT-Center for Cybersecurity Research of Catalonia
Tarragona, Catalonia
cristina.perez@urv.cat

Alberto Blanco-Justicia
Universitat Rovira i Virgili
Dept. of Computer Science and Mathematics
UNESCO Chair in Data Privacy
CYBERCAT-Center for Cybersecurity Research of Catalonia
Tarragona, Catalonia
alberto.blanco@urv.cat

ABSTRACT

An ever increasing number of decisions affecting our lives are made by algorithms. For this reason, algorithmic transparency is becoming a pressing need: automated decisions should be explainable and unbiased. A straightforward solution is to make the decision algorithms open-source, so that everyone can verify them and reproduce their outcome. However, in many situations, the source code or the training data of algorithms cannot be published for industrial or intellectual property reasons, as they are the result of long and costly experience (*e.g.* this is typically the case in banking or insurance). We present an approach whereby individual subjects on whom automated decisions are made can elicit in a *collaborative* and *privacy-preserving* manner a rule-based approximation of the model underlying the decision algorithm, based on limited interaction with the algorithm or even only on how they have been classified. Furthermore, being rule-based, the approximation thus obtained can be used to detect potential discrimination. We present empirical work to demonstrate the practicality of our ideas.

CCS CONCEPTS

• **Software and its engineering** → **Empirical software validation**; • **Information systems** → *Collaborative and social computing systems and tools*; Data analytics; • **Social and professional topics** → *Automation*; *Codes of ethics*; *Technology audits*; • **Security and privacy** → Privacy-preserving protocols; • **Computing methodologies** → Machine learning.

KEYWORDS

Machine Learning, Transparency, Explainability, Auditing, Privacy

ACM Reference Format:
Josep Domingo-Ferrer, Cristina Pérez-Solà, and Alberto Blanco-Justicia. 2019. Collaborative Explanation of Deep Models with Limited Interaction for Trade Secret and Privacy Preservation. In *Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web (FATES'19), May 2019, San Francisco, California USA.* ACM, New York, NY, USA, 7 pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
FATES'19, May 2019, San Francisco, California USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6675-5/19/05…$15.00

1 INTRODUCTION

Recent advances in machine learning and artificial intelligence, tightly coupled with the great availability of data, are driving companies towards automated decisions. Computers can make decisions much faster than their human counterparts while taking many more variables into account. Artificial neural networks and support vector machines are prevalent machine learning techniques in current decision-making tools, which results in a major drawback: the inability or at any rate the difficulty to explain why a certain decision was made on certain input data. Explaining machine learning models is of great interest both for technical, legal and ethical reasons. On the technical side, explainable machine learning allows data scientists to fine-tune their models to achieve better performance. On the legal and ethical sides, new legislation such as the EU General Data Protection Regulation [4] and recommendations like the European Commission's Ethics Guidelines for Trustworthy AI [9] urge organizations making automated decisions to be ready to explain them at the request of the affected people.

A data scientist can certainly get some insight on a particular decision by looking at the activations of a neural network, or even about the whole model by visualizing the weights in convolutional layers, but this information is usually of little to no interest to the layman. Older machine learning models based on rules were much more expressive in this regard, because just looking at the rules triggered to reach a certain decision was an explanation in itself. However, rule-based models have been clearly outperformed by (the less explainable) deep models.

Several works, such as [13–16, 19], propose mechanisms to equip deep models with explainability. A usual approach is to treat the deep model as a black box and use the decisions of the model to build simpler or more localized models that are easier to explain. These explainability proposals, however, make assumptions that limit their use, again, to data scientists rather than the general public. Typical assumptions are knowledge of the whole training data set or unrestricted access to the deep models (even if these are treated as black boxes). An individual who has been affected by an automated decision is very unlikely to have access to the training data or to the deep model:

- On the one hand, deep models require large amounts of data to be trained, which normally come either from an already existing service by the same organization offering the machine learning service or from a data market. Thus, in most cases these data sets cannot be made available to the public because, first, they are a highly valued asset of the organization that has collected the data and, second, they may contain personally identifiable information, and thus be subject to privacy laws.
- On the other hand, the training process of deep models needs to be carefully fine-tuned, from the architecture of the model to meta-parameters. This fine tuning takes time and computing power (and therefore money), which makes a well-trained deep model another highly valued asset that organizations most probably wish to protect as a trade secret.

Beyond explainability, non-discrimination is also a societal requirement on automated decisions. Decision-making tools should not discriminate people belonging to disadvantaged or vulnerable groups as established in the Universal Declaration of Human Rights. Therefore, deep models should not produce biased decisions based on gender, ethnicity, sexual orientation, religious beliefs, etc. If decisions can be explained in terms of rules, then anti-discrimination frameworks such as those proposed by Hajian *et al.* [6–8] can be used to detect discriminatory rules and sanitize the training data set so that the resulting rules be non-discriminatory.

Contribution and plan of this paper

In this paper, we present a methodology that allows individual subjects on whom automated decisions have been made to elicit in a collaborative and privacy-preserving manner a rule-based approximation of the model underlying the decision algorithm. Individuals can do so based on limited interaction with the algorithm or even only on how they have been classified. Furthermore, being rule-based, the approximation thus obtained allows detecting potential discrimination. Overall, the proposed system provides a way to empower users against opaque decision-making systems. Therefore, users are incentivized to collaborate for their own benefit, which in turn benefits the entire users' community: this is known as a co-utile situation [2].

We base our proposed methodology on building explanatory rule-based models for specific queries to a deep model. While this is not entirely new, the novelty of our proposal is that we assume that service providers limit the number of queries a single user can make, in order to preserve their deep model as a trade secret. This limitation makes it impractical for a single user to probe the model. To overcome this problem, we propose two collaborative methods:

- One method is predicated on the availability of a simulator of the deep model (such as those offered by insurance companies to potential customers to calculate fees based on personal details). Each user can query this simulator a limited number of times. By agreeing on the queries to be made by each user, a collaborating community of users can cover the entire domain of input attributes and thereby build a general explanatory approximation of the deep model. Sharing the input features and output labels of the queries does not pose privacy problems because in general users do not query with their own true input features.
- On the other hand, we explore a more "frugal" method for the case in which no simulator is available and each user is only allowed to query the deep model once, in principle with her own features. To build the model approximation, users need to share their input features and the labels returned by their queries. This poses privacy problems if users share their true features and accuracy problems if they do not. We propose solutions to deal with this privacy-accuracy trade-off.

Section 2 gives background on explainability, anti-discrimination sanitization and the conflict with trade secret. Section 3 describes the collaborative rule-based model approximation method with a simulator and Section 4 describes the method without a simulator. Section 5 presents empirical results on the performance of both methods. Finally, conclusions and future research lines are gathered in Section 6.

2 BACKGROUND

2.1 Explainability in machine learning

Article 22 of the General Data Protection Regulation (GDPR, [4]) states that citizens may not be subjected to decisions based only on automated mechanisms, and may request an explanation of such decisions [5]. It has however been noted in [21] that the explainability requirement in the GDPR is not clear enough. Nonetheless, other documents by the European Commission, such as the Ethics Guidelines for Trustworthy AI [9], distinctly emphasize the need for explainability. Unfortunately, explaining currently employed machine learning models is a highly complicated matter, since their complexity makes it extremely difficult to interpret their decisions.

The arduousness of explaining deep models is unsettling not only for the individuals affected by their decisions, but also for the artificial intelligence practitioners themselves. If a model makes mistakes, being able to interpret its wrong decisions would greatly help AI engineers to tune certain parameters and/or improve the curation and processing of the training data sets.

Several methods have been proposed to generate explanations for machine learning algorithms whose internal details are undisclosed. Some of these methods are focused on explaining decisions made by a particular model, *e.g.* a neural network [16] or a support vector machine [13], whereas others are model-agnostic [14, 15, 19] and allow explaining decisions made by any classification algorithm without being aware of its underlying model.

Machine learning explainability methods can also be classified according to the explanatory models they use. Some proposals provide visual and textual explanations, like [10]. Yet, the most common explanatory models are decision rules [15, 19], decision trees [17], and linear models [18]. For any of these explanatory models to be useful for human interpretation, the complexity of the explanations must be low enough. For instance, a two-level decision tree may be interpretable for humans, whereas a hundred-level decision tree is unlikely to serve the purpose.

Yet another categorization is between local and global explanatory methods. Some solutions generate local explanatory models that approximate the hidden (deep) model in the vicinity of a given point [13–15, 19]. In contrast, other solutions aim at explaining the

entire model [20]. In either case, the goal is for the explanatory models to predict the outcome of the hidden models as accurately as possible.

We next review in slightly more detail two of the above-mentioned proposals, that are model-agnostic and local. LIME [14] provides local explanations of decisions by sampling instances around the provided feature set (*i.e.* the user query) and interrogating the model to obtain the corresponding labels. This set is then used to train an explainable local model, such as decision trees. The authors test whether the explanations are faithful to the original model, whether the explanations increase the trust in the predictions and the model, and whether human non-experts can choose the best models and improve them. The same authors later proposed a method dubbed Anchors [15], in which local explanations of predictions are provided as rules (called anchors). An anchor is defined as a decision rule that has as precision (with respect to the hidden model) over a threshold. For a given prediction, different anchors may exist, and the authors propose algorithms to search for the one with the highest coverage, that is, the one that matches the largest amount of input samples. The paper also provides experimental evaluation of the ability of anchors to predict new unseen instances, both in terms of the amount of predictions that can be performed with a set of anchors computed from other inputs and the precision of those predictions with respect to the hidden model.

2.2 Discrimination detection and correction

Being able to explain decision results is a major asset to identify and correct discrimination or bias in machine learning models. While an automated classifier may be regarded as a fair decision-making tool, it is not necessarily so. In particular, if the classifier was trained on inherently biased data, the model it learned is likely to result in discriminatory decisions. Works by Hajian *et al.* [6–8] use rule-based models to tackle discrimination detection and correction, either by identifying and pre-processing inherently discriminatory instances in the training data, or by directly acting on the mined rules by eliminating some of them and/or generalizing some of their conditions. Clearly, these strategies are intended for explainable models, because changing the mined rules in a utility-preserving way requires understanding them. In [23], the authors focus on a different kind of discrimination, namely the different misclassification rates in different groups of people. The authors argue that a classifier should perform similarly for all instances, and propose a methodology that requires access to the ground truth (*i.e.* the training data).

2.3 Conflict between explainability, anti-discrimination and trade secret

All the above approaches to explainability and anti-discrimination either assume complete knowledge of the training data set or unlimited access by the users to the decision-making tool (although [15] explicitly tries to limit the amount of queries). We argue that these assumptions are unrealistic in production environments where the user is the subject affected by the decisions, rather the designer of the decision-making tool. The reason is that, as discussed in Section 1, both the training data and the decision model are usually protected by the service provider as trade secrets.

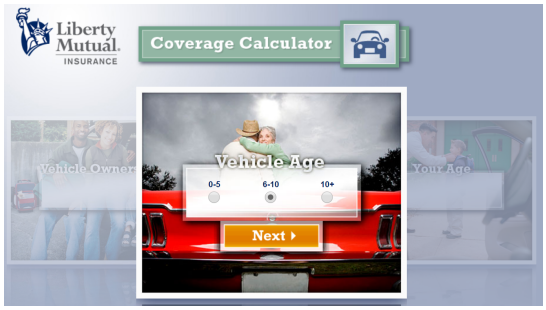


Figure 1: Simulator to calculate car insurance premiums

3 COLLABORATIVE RULE-BASED MODEL APPROXIMATION WITH A SIMULATOR

Although the internals of corporate decision-making algorithms are often hidden to the public, in some scenarios the service providers publish tools that allow querying their models *in a limited or controlled way*. The internals of the model remain hidden, but users are able to obtain responses from the algorithm for *a few* feature sets. Take as examples the web-based simulators for mortgage loans or car insurances that some banks or insurers make available on their websites[1]. Prospective customers can fill in their data and get a quotation for the monthly cost of a mortgage or the premium of an insurance. Figure 1 shows one such simulator.

Using public simulators to query a hidden model has one immediate consequence. Since users fill in the input data for the simulator themselves, the inputs may not correspond to the real user data (users can modify their inputs to observe how predictions change, or they can even lie). Therefore, they can query the model for feature sets that do not necessarily describe them. This is usually not possible whenever the evaluation is performed directly by the service provider. For instance, bank employees request proofs of income, work contracts, or official employment records in order to extract the feature sets to be used to make decisions on mortgages. Car insurance agents request the prospective customer's driving license and check previous insurance policies and reported accidents in their databases before offering a premium quotation.

To protect their model against deceptive users probing it, service providers often restrict the number of queries that any single user can submit to the simulator. This restriction is enforced, for example, by requiring the user to report a valid phone number, that is verified by the service provider before offering any results on the simulation. The simulator can also enforce throttling mechanisms in their APIs, which will limit he number of queries that any single user, or all users in general, can perform per unit of time. Other limiting strategies include the use of *CAPTCHAs* or any other anti-bot mechanism.

In this scenario, where a limited number of queries on arbitrary feature sets can be submitted by each user to a public simulator, user collaboration is a natural strategy to approximate the hidden model and understand the decisions it is making. This brings many benefits to the users and to society: (1) discriminatory behavior

[1] https://welcome.libertymutual.com/insurance-coverage-calculator-estimator/auto.html

can be detected, (2) inconsistencies in the predictions given by the model can be discovered, and (3) users unable to query the model themselves can learn the answer the model would output for their feature set.

3.1 Decision rule elicitation

Since users can query the simulator to obtain the answers to some feature sets, the most immediate technique they can use to collaboratively approximate the hidden deep model is to publish those answers. However, a more concise representation of the simulator's answers is needed to obtain human-understandable explanations of the hidden deep model.

Therefore, in addition to publishing the individual predictions each user learns from the simulator, we propose to mine *decision rules* from the pooled predictions. These rules have the advantage of explaining the model predictions and, at the same time, they are concise and they describe the predictions of the simulator for many possible feature sets.

A decision rule is an expression $X \rightarrow C$, where X is a a set of features (the premise or body of the rule) and C is a classification label. The *support* of the set of features X, $supp(X)$ is the number of predictions that take X as part of their input. Similarly, the support of a rule, $supp(X, C)$, (also sometimes called the coverage of the rule) is the number of predictions that contain X and C. The *confidence* of a decision rule, $conf(X \rightarrow C)$, measures how often C appears in predictions that contain X; hence, if $supp(X) > 0$, we have $conf(X \rightarrow C) = supp(X, C)/supp(X)$. Finally, a *frequent* decision rule is one with a support greater than a certain threshold τ_s and with a confidence greater than a certain threshold τ_c.

We propose to use a collaborative procedure generalizing the approach of Anchors [15] for a community of users to elicit decision rules that explain the predictions they receive. Specifically, users follow Protocol 1:

Protocol 1 (Rule elicitation with a simulator).

(1) *Each user in the community does:*
   (a) *If no predictions have been published so far by other users in the community, choose a region of feature sets, not necessarily containing the user's true features (*e.g. *if features are Age and Zipcode, choose an age range and a zipcode range). On the contrary, if some predictions have been already published, choose a feature set region that contains no published predictions.*
   (b) *Query the simulator as many times as allowed by randomly picking feature sets in the chosen region. As a result, a prediction for each feature set will be obtained.*
   (c) *Publish the predictions, where each prediction includes the input feature set and the simulator's answer.*
(2) *Any user can do:*
   (a) *Mine frequent decision rules from the published predictions.*
   (b) *Publish the mined decision rules.*

As mentioned in Section 1, a user incurs no privacy risks when sharing the predictions she has obtained from the simulator, because in general the input feature sets are not hers.

3.2 Detecting discrimation

Previous works have focused on providing metrics for quantifying the degree of discrimination of a decision rule [6, 11, 12]. A commonly employed metric is the extended lift (*elift*), that quantifies the degree of direct discrimination of a rule $A, B \rightarrow C$ whose premise contains a subset of features A defining a protected/vulnerable group (*e.g.* A could contain `Gender=woman` and/or `Religion=muslim`) along with other features B. Then

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)}.$$

The intuition behind *elift* is to measure the effect of belonging to the protected group on the decision. There is no discrimination if and only if *elift* is 1, in which case belonging to the protected group is irrelevant to the decision.

The predictions and rules published as a result of Protocol 1 can be used to detect discriminatory behaviors that may exist in the hidden model. This can be done by *any user in the community* by running Algorithm 1, where we consider that a rule is discriminatory if its *elift* is greater than a threshold α:

Algorithm 1 (Detection of discrimination(α)).

(1) *For each published rule $A, B \rightarrow C$ whose premise contains a subset of features A defining a protected group, use the published predictions to compute $elift(A, B \rightarrow C)$.*
(2) *Return as discriminatory the rules whose elift is greater than α.*

If Algorithm 1 returns a non-empty set of discriminatory rules, the user running the algorithm can complain to the service provider, who should endeavor to correct that discrimination using the techniques recalled in Section 2.2.

3.3 Malicious actors

The system formed by Protocol 1 and Algorithm 1 can be easily extended to account for malicious actors. On the one hand, a malicious user may want to publish fake rules, that do not describe the behavior of the target hidden deep model, with the goal of undermining the credibility of the service provider. On the other hand, a malicious service provider may be hiding some of the features that are being used by the algorithm, and thus may generate different predictions for what seems to be the same data point from the users' perspective.

Although it is difficult to prevent both types of malicious behavior altogether, the following two minor modifications to Protocol 1 allow detecting cheaters:

- At Step 1b of Protocol 1, we require the simulator to return its responses digitally signed by the service provider. Each signed response includes the input feature set and the label returned by the simulator.
- At Step 1c the user publishes the predictions signed by the service provider.

With the above modifications, if the service provider's simulator returns two different predictions for the same feature set, there exists a proof of its misbehavior, as both predictions bear the service provider's signature. Also, users cannot inject fake predictions, because they would not carry the service provider's signature.

On the other hand, the process of mining decision rules is reproducible so that any fake rules published by malicious users can be falsified by anyone based on the signed predictions. In this way, fake rules can be discarded.

The downside of making the modifications described in this section is that they imply an additional overhead, both in terms of computational cost (generating and validating the signatures) and storage (each published prediction is appended a digital signature).

4 COLLABORATIVE RULE-BASED MODEL APPROXIMATION WITHOUT A SIMULATOR

In this scenario, the service provider does not supply any particular tool to test the deep model and obtain local explanations, so all a particular user can do is to query the deep model once, presumably to obtain the prediction/decision on her features.

If a community of users consists of individuals whose features are diverse enough, users could pool the predictions received by each of them and then engage in rule elicitation in the way described in Step 2 of Protocol 1. That would yield a set of rules offering an approximate explanation of the hidden model. These rules could also be examined to detect discrimination as per Algorithm 1.

The problem with the above approach is that if users share the predictions they receive based on their true features, they are revealing personal information, whereas if they use completely fake features a utility loss is likely. Hence, to be able to gather the large numbers of predictions required to approximate the model, we need to alleviate the privacy leakage in a clever way. We will resort to randomized response (RR, [1, 22]) to enable users to "lie" about the value of their features and/or classification label —that is, a user will be able to plausibly deny that the values she has reported are her true values— but do so in a "controlled" way —that is, in a way that still allows reconstructing the *true joint distribution* of features and classification labels.

Consider $\mathbf{X}$ to be a feature attribute that can take r possible values. Then the randomized response $\mathbf{Y}$ reported by the user instead of $\mathbf{X}$ is computed using

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \vdots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix}$$

where $p_{uv} = \Pr(\mathbf{Y} = v|\mathbf{X} = u)$, for $u, v \in \{1, \ldots, r\}$ denotes the probability that the randomized response is v when the respondent's true attribute value is u.

A strong point of RR is that, while it allows plausible deniability to users, it also allows estimating the proportion of true responses based on the reported responses. Let $\pi_1, \ldots, \pi_r$ be the proportions of users whose true values fall in each of the r categories of $\mathbf{X}$; let $\lambda_v = \sum_{u=1}^{r} p_{uv}\pi_u$ for $v = 1, \ldots, r$, be the probability of the reported value $\mathbf{Y}$ being v. If we define $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ and $\pi = (\pi_1, \ldots, \pi_r)^T$, then $\lambda = \mathbf{P}^T\pi$. Furthermore, if $\hat{\lambda}$ is the vector of sample proportions corresponding to λ and $\mathbf{P}$ is nonsingular, it is proven in [1] that an unbiased estimator of π can be obtained as

$$\hat{\pi} = (\mathbf{P}^T)^{-1}\hat{\lambda}. \quad (1)$$

However, we are not interested in Expression (1) estimating the distribution of a single feature. Rather, we want that expression to estimate the true joint empirical distribution of all features and the classification label. From that joint empirical distribution, it will be possible for users to elicit rules approximately explaining the model. As pointed out in [3], to estimate the joint distribution, we need to consider as $\mathbf{X}$ the *Cartesian product of all features attributes and the classification label*. On the other hand, the probability matrix $\mathbf{P}$ should be agreed upon by all the community of users. Thus we have the following rule elicitation protocol:

Protocol 2 (Rule elicitation without a simulator ($\mathbf{P}$)).

(1) *Each user in the community does:*
   (a) *Query the service provider's model with her own features.*
   (b) *Use RR with matrix $\mathbf{P}$ to randomize the vector of her features and the classification label.*
   (c) *Publish her randomized vector.*
(2) *Any user can do:*
   (a) *Compute the empirical joint distribution of the collection of published randomized prediction vectors.*
   (b) *Use Expression (1) on the computed empirical joint distribution to estimate the true empirical joint distribution of the features and the classification label.*
   (c) *Re-create a synthetic data set by repeating each combination of features and the classification label as many times as indicated by the joint distribution.*
   (d) *Mine frequent decision rules from the synthetic data set.*
   (e) *Publish the mined decision rules.*

5 EMPIRICAL RESULTS

In this section we present experiments based on the Adult dataset. We view the data set as the result of collaborative users pooling the responses they have obtained from a black-box model after submitting their individual queries.

5.1 Results of the method with a simulator

We replicated the experiments of the Anchors paper [15] for the Adult data set using a neural network as the hidden model. Adult contains several demographic feature attributes and a classification attribute, predicting whether an individual makes more or less than 50K$/year. That is, we trained a neural network with 80% of the samples in the Adult data set and we used the rest of the data set as test samples to compare the trained model and its rule-based approximate explanation. Specifically, we created rules to explain the predictions of the model for half of the test samples. This would correspond to the rules that users elicit by querying the simulator repeatedly and then publish to approximate the hidden deep model. Then, we used the second half of the test samples as validation data set to test how good were the extracted rules at predicting the behavior of the model. This would correspond to feature sets of other users that do not query the model themselves, but rather use the previously published rule-based approximation to obtain an estimate of the way the hidden model would classify them.

Results on rule support/coverage and precision are already reported in the original Anchors paper, and are thus omitted here. Rather, we analyze conflicting rules, that is, how often two different

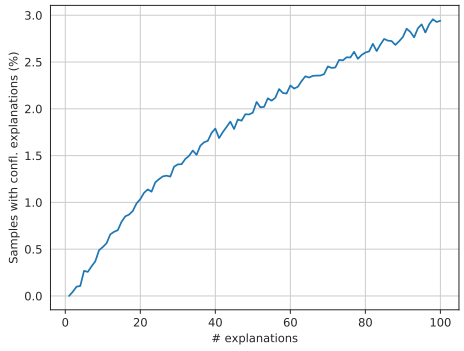


Figure 2: Percentage of samples with conflicting rules

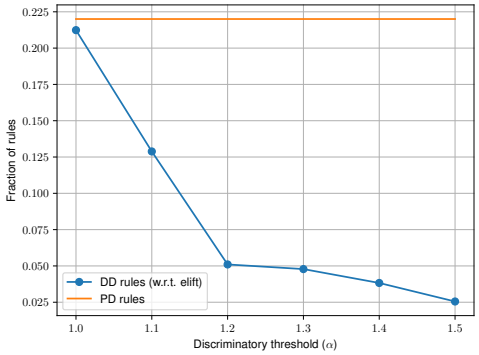


Figure 3: Fraction of potentially discriminatory rules (PD) and α-discriminatory rules (DD) with respect to $elift$.

rules are applicable to the same sample but provide contradictory predictions. Figure 2 shows the percentage of samples in the validation data set for which conflicting rules exist, as the number of published rules increases.

We have also analyzed how many of the elicited rules are discriminatory with respect to the *elift* metric. We have considered the subset of features defining a protected group to be A = {Sex = Female, Age $\leq$ 28}. There are 22% potentially discriminatory rules, that is, rules whose premises include the features in A. As shown in Figure 3, the percentage of actually discriminatory rules depends on the threshold α we take for *elift*, and ranges from 21.2% (for $\alpha = 1$) to 2.5% (for $\alpha = 1.5$).

5.2 Results of the method with no simulator

We also conducted an experiment to assess the effect of RR on the accuracy of the rules mined from the randomized predictions.

We want to test if, after using RR to protect some values from these responses, the resulting trained explainable model is close to a model trained on the raw, unaltered responses.

For that purpose we performed RR on 80% of the Adult data set for the Cartesian product of attributes *Sex*, *Education* and *Race*, so that the probability matrix had one row and one column for each combination of the three attributes.

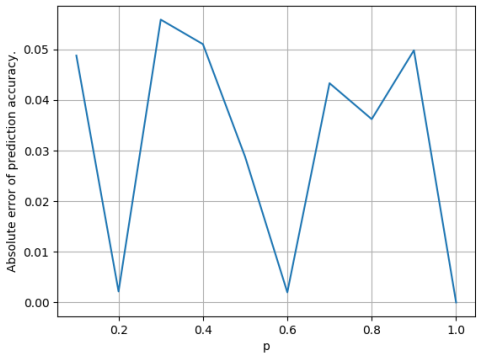


Figure 4: Prediction error (%) with rules mined from data sets randomized with probability p of leaving values unaltered

Specifically, we obtained 10 randomized versions of Adult, each with a different probability matrix $\mathbf{P}_p$, for $p = 0.1, 0.2, 0.3, \cdots, 1.0$. Matrix $\mathbf{P}_p$ had all the probabilities in its diagonal set to p and the off-diagonal probabilities set as described in Section 6.2 of [3], that is,

$$p_{uv} = (1-p)\frac{d_{uv}}{\sum_k d_{uk}},$$

where d_{uv} is the inverse of the number of differing attribute values between u and v, and $\sum_k d_{uk}$ is the sum of all changes between feature set u and all other feature sets k.

After that, we computed the estimated distributions of the data set as per Equation (1) and reconstructed the data sets by sampling random feature sets from the obtained distributions. Note that the estimated distributions can be computed from the randomized data sets and the matrices $\mathbf{P}_p$. From each reconstructed data set, we followed the substeps of Step 2 of Protocol 2, that yielded a set of mined rules. So we got 10 sets of mined rules for $p = 0.1, \cdots, 1.0$. Note that for $p = 1.0$ there was no real randomization and what was obtained were the rules mined on the original Adult data set, which we used as a baseline.

Then we evaluated the 9 sets of rules corresponding to $p = 0.1, \cdots, 0.9$ by computing their predictions (>50K, ≤50K) on the features of the remaining 20% of the Adult dataset and comparing them against the predictions given by the rules for $p = 1.0$. Figure 4 shows the percentage of Adult samples in which the 9 sets of rules gave predictions different from those of the baseline set of rules. Although the error may seem to vary a lot with p, in reality it varies within a very narrow range (0% to 5.5%).

6 CONCLUSIONS AND FUTURE WORK

We have presented a methodology that enables individuals to audit deep models that make decisions on them. The novelty of our approach is that it works even if the service provider owning the deep model restricts access to it in order to preserve its trade secret. To circumvent access restrictions, we adopt a collaborative approach. We consider two scenarios, with and without a model simulator.

When there is no simulator, we propose to use randomized response to allow users to share information without renouncing their privacy.

Future work will include testing on more data sets, fine tuning the parameters of rule elicitation and randomized response, and increasing the number of utility metrics tried. We will also deal with the dimensionality problem that arises in randomized response of Cartesian products when there are many features involved in decisions. Finally, implementing the proposed collaborative protocols with a user-friendly interface would go a long way towards effectively empowering the general public with audit capabilities.

ACKNOWLEDGMENTS AND DISCLAIMER

The following funding sources are gratefully acknowledged: European Commission (project H2020 700540 "CANVAS"), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant no. 2017 SGR 705), Spanish Government (project TIN2014-57364-C2-R "SmartGlacis"). The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and do not necessarily reflect those of UNESCO.

REFERENCES

[1] A. Chaudhuri and R. Mukerjee. *Randomized Response: Theory and Techniques*. Marcel Dekker, 1988.

[2] J. Domingo-Ferrer, S. Martínez, D. Sánchez and J. Soria-Comas. Co-utility: self-enforcing protocols for the mutual benefit of participants. *Engineering Applications of Artificial Intelligence*, 59:148-158, 2017.

[3] J. Domingo-Ferrer, R. Mulero-Vellido, and J. Soria-Comas. Multiparty computation with statistical input confidentiality via randomized response. In *Privacy in Statistical Databases-PSD 2018*, pp. 175–186. Springer, 2018.

[4] *General Data Protection Regulation*. Regulation (EU) 2016/679. https://gdpr-info.eu

[5] S. Greengard. Weighing the impact of GDPR. *Communications of the ACM*, 61(11):16–18, 2018.

[6] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459, 2013.

[7] S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 28(5-6):1158–1188, 2014.

[8] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery*, 29(6):1733–1782, 2015.

[9] The European Comission's High-Level Expert Group on Artificial Intelligence. *Draft Ethics Guidelines for Trustworthy AI*. December 2018.

[10] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*, 2016.

[11] D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 581–592. SIAM, 2009.

[12] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 560–568. ACM, 2008.

[13] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, B. Pearcy, C. MacDonell, and J. Anvik. Visual explanation of evidence with additive classifiers. In *21st AAAI Conf. on Artificial Intelligence–AAAI'06*, pp. 1822-1829. AAAI, 2006.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *32nd AAAI Conf. on Artificial Intelligence–AAAI'18* pp. 1527-1535. AAAI, 2018.

[16] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

[17] S. Singh, M. T. Ribeiro, and C. Guestrin. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*, 2016.

[18] E. Strumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18, 2010.

[19] R. Turner. A model explanation system. In *IEEE Intl. Workshop on Machine Learning for Signal Processing–MLSP'16*. IEEE, 2016.

[20] M. M. C. Vidovic, N. Görnitz, K.-R. Müller, and M. Kloft. Feature importance measure for non-linear learning algorithms. *arXiv preprint arXiv:1611.07567*, 2016.

[21] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[22] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[23] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proc. of 26th Intl. Conf. on World Wide Web*, pp. 1171–1180. 2017.

Privacy-aware Linked Widgets

Javier D. Fernández
Vienna University of Economics and Business
Complexity Science Hub Vienna
javier.fernandez@wu.ac.at

Fajar J. Ekaputra
Vienna University of Technology
fajar.ekaputra@tuwien.ac.at

Peb Ruswono Aryan
Vienna University of Technology
peb.aryan@tuwien.ac.at

Elmar Kiesling
Vienna University of Technology
elmar.kiesling@tuwien.ac.at

Amr Azzam
Vienna University of Economics and Business
amr.azzam@wu.ac.at

ABSTRACT

The European General Data Protection Regulation (GDPR) brings new challenges for companies, who must demonstrate that their systems and business processes comply with usage constraints specified by data subjects. However, due to the lack of standards, tools, and best practices, many organizations struggle to adapt their infrastructure and processes to ensure and demonstrate that all data processing is in compliance with users' given consent. The SPECIAL EU H2020 project has developed vocabularies that can formally describe data subjects' given consent as well as methods that use this description to automatically determine whether processing of the data according to a given policy is compliant with the given consent. Whereas this makes it possible to determine whether processing was compliant or not, integration of the approach into existing line of business applications and ex-ante compliance checking remains an open challenge. In this short paper, we demonstrate how the SPECIAL consent and compliance framework can be integrated into Linked Widgets, a mashup platform, in order to support privacy-aware ad-hoc integration of personal data. The resulting environment makes it possible to create data integration and processing workflows out of components that inherently respect usage policies of the data that is being processed and are able to demonstrate compliance. We provide an overview of the necessary meta data and orchestration towards a privacy-aware linked data mashup platform that automatically respects subjects' given consents. The evaluation results show the potential of our approach for ex-ante usage policy compliance checking within the Linked Widgets Platforms and beyond.

CCS CONCEPTS

• **Security and privacy → Information accountability and usage control**; **Privacy protections**.

KEYWORDS

Privacy; GDPR; Compliance; Linked Data

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317591

ACM Reference Format:
Javier D. Fernández, Fajar J. Ekaputra, Peb Ruswono Aryan, Elmar Kiesling, and Amr Azzam. 2019. Privacy-aware Linked Widgets. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317591

1 INTRODUCTION

The recent European General Data Protection Regulation (GDPR) [1] defines a set of obligations for controllers and processors of personal data. Among other requirements, companies must be transparent about their processing of personal data and about its sharing within and between organizations. Furthermore, companies also need to demonstrate that their systems and business processes comply with usage constraints specified by data subjects.

Thus, the GDPR implicitly fosters requirements for more accountable systems that are aware of restrictions on personal data that is being collected, used, and shared. Designing and implementing a GDPR-compliant infrastructure that fosters privacy-by-design, is, however, a challenge. Several tools [9, 13, 14] aim to assist companies in assessing their GDPR compliance. These tools are limited in that they focus on privacy impact assessment based on standard questionnaires, but they cannot be used to automatically check compliance with subjects' policies.

Other initiatives [2, 16] use semantic web technologies to represent policies in a manner so that they can be checked automatically. In this context, the EU H2020 SPECIAL[1] project provides (i) a GDPR-oriented policy language for subject's consent, (ii) vocabularies to represent data processing events, and (iii) a complete architecture [11] for GDPR transparency and compliance.

In this paper, we focus on building *ex-ante* compliance mechanisms into data processing infrastructures. This is motivated by the need to ensure that any potential violations of subjects' restrictions on the use of their data are detected before any infringing processing occurs. To this end, we propose an approach to integrate the SPECIAL policy language and compliance checking algorithm [2] into a mashup environment – the Linked Widgets Platform (LWP) – [18] and its constituent components.

LWP is a framework for defining Linked Data-based components (e.g. data ingestion, semantification, search, integration, analysis, etc.) and to interconnect them in a processing workflow that can

[1] https://www.specialprivacy.eu/

consume and manage data from heterogeneous sources, making use of the flexibility and expressivity of semantic technologies. The privacy-aware LWP extension, consisting of SPECIAL-based policy metadata and compliance orchestration, automatically checks that a LWP workflow is compliant with the subject's policies of the data being processed, preventing non-compliant workflows from being executed. Our initial setup and experiments with realistic data and policies in the smart energy domain, outlined in Section 3, have yielded encouraging results and demonstrate the feasibility of the ex-ante (i.e. at runtime) GDPR checking approach and the applicability of SPECIAL in existing Line of Business systems.

The remainder of the paper is organized as follows. Section 2 provides background on the SPECIAL framework and the LWP. Section 3 then presents our proposal towards a privacy-aware LWP, integrating the SPECIAL components to provide ex-ante GDPR compliance of the LWP workflows. Section 4 evaluates the privacy-aware LWP prototype in our case study, focused in the smart energy domain. Finally, Section 5 summarizes the state of the art, and Section 6 concludes and devises future work.

2 BACKGROUND

In this section we briefly review the SPECIAL usage policy language and compliance mechanism and provide an overview of the Linked Widget Platform, which will be extended in the next section.

2.1 SPECIAL Transparency and Compliance

The SPECIAL consent, transparency and compliance framework consists of two primary components, (i) the SPECIAL Consent component that uses the SPECIAL policy language to represent the consent from the data subject in the form of a usage policy; and the (ii) the SPECIAL Transparency and Compliance Component that uses the SPECIAL log vocabulary to represent data processing events, and the SPECIAL compliance checking mechanisms to verify compliance of such events with the usage policies. In the following we briefly present the SPECIAL usage policy and the compliance checking mechanisms. The log vocabulary[2] follows on from the usage policy and is out of the scope of this paper.

The SPECIAL Policy Language. Conceptually, a *usage policy* is meant to specify a *set of authorized operations*. According to the GDPR, these policies shall specify clearly (i) which data are collected, (ii) what is the purpose of the collection, (iii) what processing will be performed, (iv) where and for how long is the storage of the data, and (v) whether or not the data will be shared with others. The SPECIAL policy language follows these five principles and represent them using semantic technologies. Thus, a SPECIAL usage policy, P_s, is composed of one or more *basic usage policies*, each of which is an OWL 2 [7] expression of the form:

```
ObjectIntersectionOf(
 ObjectSomeValuesFrom(spl:hasData SomeDataCategory)
 ObjectSomeValuesFrom(spl:hasProcessing SomeProcessing)
 ObjectSomeValuesFrom(spl:hasPurpose SomePurpose)
 ObjectSomeValuesFrom(spl:hasRecipient SomeRecipient)
 ObjectSomeValuesFrom(spl:hasStorage SomeStorage) )
```
(1)

SPECIAL provides taxonomies that represent general categories for each case, which can be extended for particular cases. For instance, Listing 1 shows an example of a SPECIAL usage policy to represent that a subject consents to collect the energy consumption, to integrate other data sources and to perform profiling on the anonymous data in order to allow the company to optimize the energy infrastructure, who can store the data in EU indefinitely. This example makes use of the aforementioned SPECIAL auxiliary vocabularies (*spl*, *svpr*, *svl*, *svdu*, *svr*), an existing extension for cyber-physical social systems[3] (*svd-cpss*), and an exemplary vocabulary potentially defined by a company (*eg*).

Listing 1: Example of a SPECIAL usage policy on energy data

```
ObjectIntersectionOf(
    ObjectSomeValueFrom( spl:hasData eg:EnergyConsumption )
    ObjectSomeValueFrom( spl:hasProcessing
        ObjectIntersectionOf(
            eg:Profiling svpr:Anonymize eg:Integration svpr:Collect ))
    ObjectSomeValueFrom( spl:hasPurpose svd-cpss:Optimizing )
    ObjectSomeValueFrom( spl:hasStorage
        ObjectIntersectionOf(
            ObjectSomeValuesFrom( spl:hasLocation svl:EU )
            DataSomeValuesFrom( spl:hasDuration svdu:Indefinitely ))
    ObjectSomeValueFrom( spl:hasRecipient svr:Ours ) )
```

Using SPECIAL for Compliance Checking. In SPECIAL, policies and log events are described in semantically unambiguous terms aligned to the same taxonomies defining usage policies, hence it facilitates transparency and automatic compliance checking. Regarding this latter, the usage policy enforced by a data controller contains the operations that are permitted within the data controller's organization. Therefore, the usage U_c attached to a SPECIAL log entry *complies* with the usage policy P_s in the data subject's consent if and only if all the authorizations in U_c are also authorized by P_s, that is, U_c complies with P_s if and only if

$$U_c \subseteq P_s. \tag{2}$$

Thus, in OWL 2 terminology, this implies checking whether the following axiom is *entailed* (implied) by the combined ontology O containing the SPECIAL policy language ontology plus the aforementioned auxiliary vocabularies:

$$\texttt{SubClassOf}(U_c\ P_s). \tag{3}$$

This is inherently supported by general inference engines for OWL 2 (e.g. HermiT and FaCT++). Further details of the compliance checking mechanism can be found in [2, 10].

For instance, a log entry can specify that there is a process of type `eg:SensorGathering` on location data. This entry is compliant with a potential usage policy stating that the controller can collect (`svpr:Collect`) location data *iff* `eg:SensorGathering` is a subclass of `svpr:Collect`.

This mechanism can be used for ex-post compliance checking, i.e. based on event logs, as well as ex-ante compliance, i.e. operations to be performed on the subject's data. In the next section, we focus on the latter, integrating it into the Linked Widgets Platform.

[2] http://purl.org/specialprivacy/splog

[3] https://w3id.org/cityspin/ontology/special-cpss/0.1.0/index-en.html

2.2 Data Mashups and Linked Widgets Platform

Mashup environments are designed to support non-expert users in combining and processing data from multiple sources to create a single new service displayed in a graphical interface [6]. In corporate settings, mashups can facilitate lightweight composition of heterogeneous enterprise applications in a shorter time to cover the long tail of user needs [8]. The term implies easy, fast integration, frequently made possible by access to open APIs and data sources to produce results beyond the predictions of the data owners [3]. This focus on ad-hoc data integration is a major strength of the mashup paradigm, but when personal data is involved, it raises significant privacy concerns. In particular, flexible and unconstrained ad-hoc integration and processing of personal data can easily clash with requirements for informed consent and transparency. Hence, it is important to (i) provide data subjects with the means to explicitly state the permissible use of their data, and (ii) make sure that a mashup platform that processes personal data is aware of and respects these usage policies. Given that mashups are typically constructed in an ad-hoc manner, it is necessary to build ex-ante conformance checking directly into the platform in order to ensure compliance at runtime.

We illustrate this by extending the *Linked Widgets Platform* [19] with privacy-aware mashups and processing components. This platform combines semantic web and mashup concepts to allow users to collaboratively and interactively integrate data in an ad-hoc and distributed manner. Each stakeholder can contribute their data and computing resources to a shared data processing flow. In addition, the platform facilitates both ad-hoc and persistent data integration on multiple devices.

Figure 1 illustrates a simple example mashup created in the Linked Widget Platform that integrates data on economic and environmental indicators from independent sources. The mashup consists of four widgets: the *Economic Statistics* and *Environmental Statistics* widgets provide data from the World Bank[4] and European Environment Agency[5] SPARQL endpoints, respectively. Based on the chosen parameters of countries and indicators, these widgets execute SPARQL queries. The results are processed and transformed into W3C Data Cube vocabulary[6] data sets using the StatSpace engine [4]. In each data set, the two dimensions are country and date; a single measure is the selected indicator. Next, the two data cubes are integrated by the *Cube Merger* widget and finally, the processed data is visualized by the *C3 Chart* widget.

3 THE PRIVACY-AWARE LWP

In this section, we present how the SPECIAL framework can be integrated into LWP to allow *ex-ante* compliance checking of usage policies. First, we provide a brief overview of the envisioned *ex-ante* compliance scenario (Section 3.1). Then, we detail the LWP SPECIAL-based metadata and orchestration to perform such scenario (Section 3.2), hence fostering the development of a privacy-aware LWP.

[4]http://worldbank.270a.info/sparql
[5]http://digital-agenda-data.eu/data/sparql
[6]https://www.w3.org/TR/vocab-data-cube/

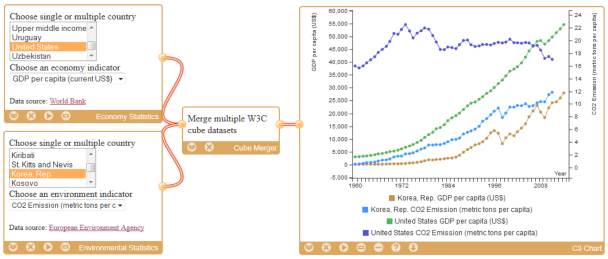


Figure 1: A mashup example on statistical data (adapted from [17])

3.1 *Ex-ante* Compliance: Scenario and Setup

Our case study is the smart (energy) building domain, where the building management (data controller) would like to analyze the energy consumption data of tenants (data subject) in relation with user behaviour as well as ambience data (e.g., outside temperature). The building management has access to the temperature data coming from building sensors, while the tenant provides data about their energy consumption as well as their personal profile.

Figure 2 provides an overview of the *ex-ante* compliance checking process. First, we assume a scenario where data subjects (e.g., a building tenant in our case study, a user of a company APP, etc.) give explicit consent to data controllers (e.g. the energy provider, or the company providing a service in an APP), according to the GDPR. We also assume that each data subjects' consent is represented as a usage policy, P_s, following the SPECIAL Policy Language (c.f. see Section 2.1). Automatic means of representing or obtaining such consents are out of the scope of this paper. Then, our scenario considers that the company (for simplicity, the same controller) wants to perform a data-intensive process (e.g. to analyze energy consumption patterns), making use of both personal data from the aforementioned data subjects and non-personal data (e.g., environmental sensor or other external sources). In this context, before any processing is performed, the company can build a "workflow" constraint, W_c, considering the SPECIAL dimensions, i.e., (i) the category of data to be processed, (ii) the purpose of the processing, (iii) which kind of processing will be performed, (iv) where and for how long is the storage of the data, and (v) whether or not the processing requires to share data with others. The workflow constraint will be then represented with the SPECIAL Policy Language, hence it can be then checked against the individual data subject's usage policy before the data processing is conducted (i.e., *ex-ante*) to make sure that only personal data compliant to the data controller policies being processed. Similarly to the *ex-post* scenario considering data logs, the *ex-ante* compliance checking process consists of a simple inference task to assure that:

$$\texttt{SubClassOf}(W_c\ P_s). \tag{4}$$

Note that *ex-ante* compliance checking can be performed a) at run-time, i.e., the first step of the data processing workflow considers an input stream of personal data and the algorithm automatically verifies that each "record" is compliant with the full workflow constraint, discarding the record otherwise, or b) as a filtering batch process, i.e., the full personal data stored is checked against the

workflow constraint, and the adequate candidates are filtered and stored separately. In the following, we are agnostic of these possibilities and we assume that usage policies are performed once per usage policy. Efficient mechanisms of performing batch compliance checking (e.g. grouping usage policies based on common hierarchical elements) is devoted to future work.

3.2 Making LWP "SPECIAL"

The flexible, semantic-based LWP framework enables an efficient integration of the presented scenario for *ex-ante* compliance checking. Figure 3 shows an schematic example considering a processing workflow composed of six tasks, each of them wrapped as a single widget: Personal Data collection, Anonymization, Integration (e.g. with building sensor data), Aggregation, Analysis (e.g. using Machine Learning) and Reporting.

The main SPECIAL extension to LWP consists of two metadata levels. First, at widget level, we semantically represent the specific processing conducted by a specific widget, and we align it with the corresponding SPECIAL vocabularies. In concrete, we specify i) the categories of the data that are processed (both as input and output of the widget), ii) the concrete process tasks and iii) the required storage (location and duration). Note that all these categories are optional, as some widgets perform general tasks or they abstract the data being processed (e.g. they do not restrict the category of data). These "semantic annotations" using the SPECIAL vocabularies are intrinsic of the widget regardless of their usage in different LWP mashups.

Then, at mashup level, i.e. for a particular processing workflow in LWP, we allow the data controller to specify the designated purpose(s) and recipient(s) of the mashup, using the corresponding SPECIAL vocabularies. Similarly, both types can be optionally present, but a more fine-grained specification enables to obtain a result for the *ex-ante* compliance checking that closely reflects the adequation to the constraints of data subjects.

Finally, the combination of widget and mashup-level metadata using the aforementioned SPECIAL vocabularies serves as the basis for constructing the data controller "workflow" constraints. It is worth mentioning that two "aggregation" methods are possible to obtain the final constraints:

- **Mashup-level constraint aggregation**. In this method, a single workflow constraint is created, representing the constraints of the entire mashup. To this end, we combine the individual categories of each widget i) applying `owl:unionOf` for the data category (i.e the mashup contains a union of the different data categories processed in the workflow), ii) using `owl:intersectionOf` for the process category (i.e. the mashup consists of a conjunction of several processes), and iii) joining the different storage periods [10]. In each category, we additionally remove classes that are sub-classes of other class.
- **Widget-level constraint aggregation**. Within this method, instead of having a single usage policy for the mashup, we provide one usage policy per widget within a mashup. To this end, we add the mashup types (purpose and recipient) to each individual widget, creating one constraint for each connected widget.

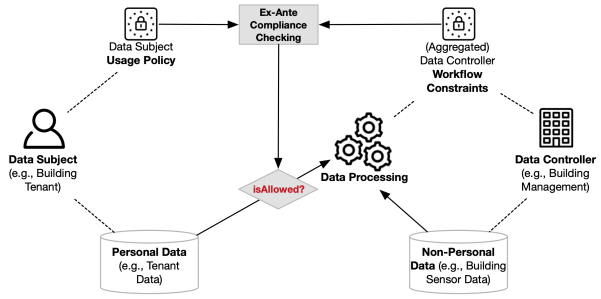


Figure 2: The overview figure of ex-ante usage-policy compliance checking

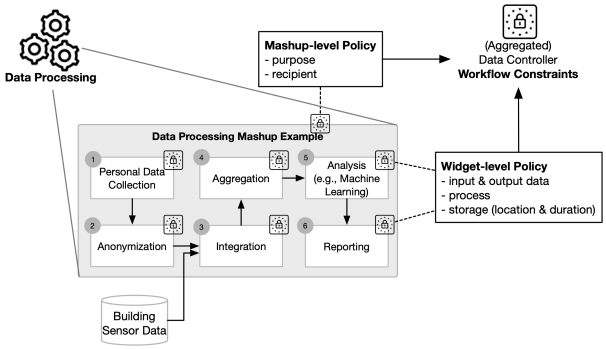


Figure 3: The components for generating data controller usage-policies

In both cases, at the end of the process, if a type is empty, we use the top level class of each type. Note that mashup-level constraint aggregation enables a direct application of the compliance checking algorithm shown above. In contrast, the algorithm should be modified to consider widget-level constraint aggregation, as special restrictions in the data subjects' policies, such as `owl:intersectionOf` relations, should be checked against different workflow constraints. Thus, in what follows, we focus on mashup-level constraint aggregations, while widget-level constraint aggregations are considered for future work.

4 EVALUATION

The LWP SPECIAL extension and the corresponding ex-ante compliance checking mechanisms have been implemented in a fully functional prototype[7]. Note that, to verify the implementation, we defined a set of criteria that a privacy-aware mashup environment should fulfill:

(1) Support ex-ante compliance checking,
(2) Conformance checks must provide correct results,
(3) Ensure that no personal data is processed without given consent,
(4) Ability to support arbitrary number of heterogeneous usage policies from various data subjects,

[7]Our privacy-aware LWP prototype and the test data are available at: https://github.com/linkeddatalab/LWP-SPECIAL

(5) Optionally, ability to identify the specific components (or lack of) leading to consent violations.

The prototypical implementation of our privacy-aware mashup platform fulfills all the mandatory requirements, while the optional identification of problematic components is considered as future work. Note also that while we programmatically ensure that all processed data fulfill subject consents, we do not focus here on other irrevocable techniques (e.g. cryptographic methods).

In the following we first evaluate our proposal on our smart energy building scenario and the workflow depicted in Figure 3. Then, we perform an scalability test on multiple workflows of different size.

All experiments run on an Ubuntu 14.05.05 Linux Server in a Intel Xeon CPU E5-2620 2GHz machine with 2GB memory.

4.1 Realistic Scenario

Following from our case study on smart energy building, we implement a LWP workflow including the components and relationships depicted in Figure 3. Table 1 shows the concrete metadata to annotate the mashup-level properties and each of the workflow components, using the SPECIAL vocabularies. Note that we split the personal data collection process into three components, which corresponds to the location, consumption data and temperature gathering processes. The last row of the table shows the final categories of the mashup-level constraint aggregation (as explained in Section 3.2).

Then, in order to test the performance of the ex-ante compliance checking on the provided workflow, we define different data subject policies. In particular, we take as input the same vocabulary used in the LWP workflow and we extend it with additional terms (for each category) in the SPECIAL auxiliary vocabularies. We then randomly generate 100 simple policies (i.e. only considering one term per category) and 100 complex policies (also considering disjunction and conjunction of terms).

The performance results of the ex-ante compliance checking, for both simple and complex policies, are shown in Figure 4. Several comments are in order. First, note that the median time to check the compliance of each policy is 23ms per policy, in both simple and complex cases. This result shows the feasibility of the system in a realistic scenario. A closer study to the difference between simple and complex policies shows that the 0.75 quantile of simple policies performs in less than 20ms, while in complex policies, the same quantile carries out the compliance checking in 18ms. This shows that the presence of very restricted policies can produce early fails in the compliance checking process and that, in general, the influence of complex policies in the performance is negligible.

4.2 Scalability Test

Our scalability test regards the ability of the approach to scale, not only to the complexity of the data subject policies, but the number of components in the workflow (i.e. the complexity of the workflow constraints). To this aim, we consider as input the same categories as our simple and complex policies, and we randomly generate and annotate random LWP workflows with 5, 10, 20, 50 and 100 components (i.e., widgets).

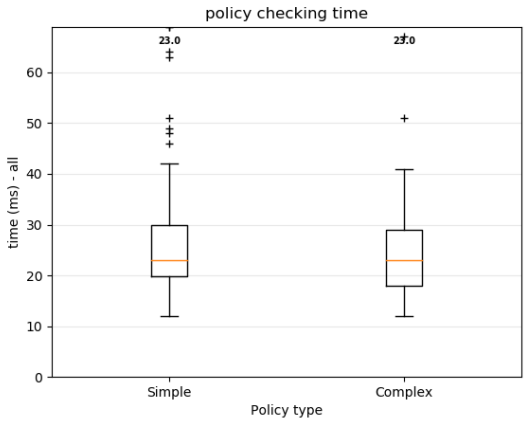


Figure 4: Results of the ex-ante compliance checking for the smart energy building workflow.

Figure 5 shows the performance results, for simple and complex policies, of random workflows at increasing number of processing components. The small variations in performance (a median of 23-27ms per policy, even in the case of 100 components) show that the proposed ex-ante compliance mechanisms in LWP is able to scale w.r.t the number of processing components in the mashup.

5 STATE OF THE ART

As for GDPR compliance, the Information Commissioner's Office (ICO) in the UK [9], Microsoft [13], and Nymity [14] have developed compliance tools that enable companies to assess the compliance of their applications and business processes by completing a predefined questionnaire. Recent works also look at the challenges of representing GDPR concepts and obligations [15, 16] as well as informed consent [5]. In contrast to existing approaches, SPECIAL proposes vocabularies [10] that can be used to record both usage policies and data processing and sharing events in a manner that supports automatic compliance checking.

Event management for business process compliance monitoring and process mining [12, 20] can be seen as orthogonal work.

6 CONCLUSIONS AND FUTURE WORK

The General Data Protection Regulation (GDPR) has brought the need of more transparent and accountable systems, assuring that any data process respects the recorded data subjects' consent. This work advances in this direction, providing an extension of an existing semantic data processing tool, the Linked Widget Platform (LWP), that automatically checks for ex-ante compliance w.r.t. the defined subject's consents. To this aim, we integrate the semantic policies and inference mechanisms provided in SPECIAL, an EU H2020 project on GDPR transparency and compliance. Our initial results on a realistic smart energy scenario, performing ex-ante compliance checking in 23ms per policy, shows the feasibility and efficiency of the SPECIAL and LWP combination, promoting the development of privacy-aware tools on the basis of semantic technologies.

Table 1: SPECIAL vocabularies for the smart energy building workflow.

Component	SPECIAL category				
	Data	Processing	Purpose	Storage	Recipient
Mashup-level	-	-	svpu:Develop	-	svr:Ours
Personal Data Collection 1	svd:Location,svd:UniqueId	svpr:Transfer	-	-	-
Personal Data Collection 2	svd:Location,cpss:ConsumptionData	svpr:Transfer	-	-	-
Personal Data Collection 3	svd:Location,cpss:Temperature	svpr:Transfer	-	-	-
Anonymization	svd:Anonymized	svpr:Anonymize	-	-	-
(Geo)Integration	svd:Location	cpss:Integration	-	-	-
Aggregation	svd:Derived	svpr:Aggregate	-	-	-
Analysis	svd:Derived	svpr:Analyse	-	-	-
Report	-	-	-	svl:EU, spl:AnyDuration	-
Mashup-level Constraint Aggregation	owl:unionOf(svd:Location, svd:UniqueId, svd:ConsumptionData, cpss:Temperature, svd:Anonymized, svd:Derived)	owl:intersectionOf(svpr:Transfer, svpr:Anonymize, cpss:Integration, svpr:Aggregate, svpr:Analyze)	svpu:Develop	svl:EU, spl:AnyDuration	svr:Ours

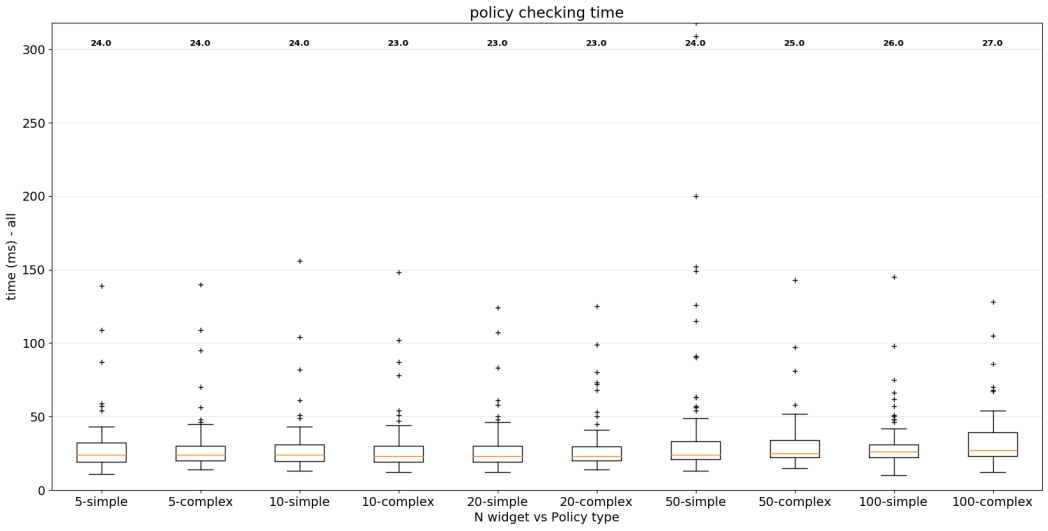


Figure 5: Results of the ex-ante compliance checking at increasing number of workflow components.

Our proposal focuses on identifying potential hazardous processes before their execution. Irrevocable means of preventing the execution is out of the scope of the paper and is subject of future work. We also plan to provide a more fine-grain report of consent violations, identifying both problematic components in the workflow or missing pieces to fulfill the constrains. Finally, we consider to apply the privacy-aware LWP in real-world environments in the more general context of smart cities.

ACKNOWLEDGEMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 731601 (SPECIAL), by the Austrian Research Promotion Agency (FFG): grant no. 861213 (CitySPIN) and 867559 (EXPEDiTE), and by the Austrian Federal Ministry for Digital, Business and Enterprise and the National Foundation for Research, Technology and Development.

REFERENCES

[1] [n. d.]. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). L119 ([n. d.]), 1–88. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

[2] Piero A Bonatti. 2018. Fast Compliance Checking in an OWL2 Fragment.. In *IJCAI*. 1746–1752.

[3] Paul de Vrieze, Lai Xu, Athman Bouguettaya, Jian Yang, and Jinjun Chen. 2011. Building enterprise mashups. *Future Generation Computer Systems* 27, 5 (2011), 637 – 642. http://www.sciencedirect.com/science/article/pii/S0167739X10001974

[4] Ba-Lam Do, Peter Wetz, Elmar Kiesling, Peb Ruswono Aryan, Tuan-Dat Trinh, and A Min Tjoa. 2016. Statspace: A unified platform for statistical data exploration. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 792–809.

[5] Kaniz Fatema, Ensar Hadziselimovic, Harshvardhan J. Pandit, Christophe Debruyne, Dave Lewis, and Declan O'Sullivan. 2017. Compliance through Informed Consent: Semantic Based Consent Permission and Data Management Model. In *Proc of PrivOn*.

[6] Darlene Fichter. 2010. What is a mashup. In *Library mashups: Exploring new ways to deliver library data*. Information Today, Incorporated.

[7] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, and Sebastian Rudolph. 2009. OWL 2 web ontology language primer. *W3C recommendation* 27, 1 (2009), 123.

[8] Volker Hoyer, Katarina Stanoevska-Slabeva, Till Janner, and Christoph Schroth. 2008. Enterprise Mashups: Design Principles towards the Long Tail of User Needs. In *2008 IEEE International Conference on Services Computing*, Vol. 2. IEEE Computer Society, Washington, 601–602. https://www.alexandria.unisg.ch/44891/

[9] Information Commissioner's Office (ICO) UK. 2017. Getting ready for the GDPR. https://ico.org.uk/for-organisations/resources-and-support/data-protection-self-assessment/getting-ready-for-the-gdpr/

[10] Sabrina Kirrane, Piero Bonatti, Javier D. Fernández, Clemente Galdi, Luigi Sauro, Daniele Dell'Erba, Iliana Petrova, and Ida Siahaan. 2018. SPECIAL Deliverable 2.8: Transparency and Compliance Algorithms V2. https://www.specialprivacy.eu/images/documents/SPECIAL_D2.8_M23_V1.0.pdf

[11] Sabrina Kirrane, Javier D. Fernández, Wouter Dullaert, Uros Milosevic, Axel Polleres, Piero A. Bonatti, Rigo Wenning, Olha Drozd, and Philip Raschke. 2018. A Scalable Consent, Transparency and Compliance Architecture. In *Proc. of ESWC*. 131–136. https://doi.org/10.1007/978-3-319-98192-5_25

[12] Linh Thao Ly, Fabrizio Maria Maggi, Marco Montali, Stefanie Rinderle-Ma, and Wil MP van der Aalst. 2015. Compliance monitoring in business processes: Functionalities, application, and tool-support. *Information systems* 54 (2015), 209–234.

[13] Microsoft Trust Center. 2017. Detailed GDPR Assessment. http://aka.ms/gdprdetailedassessment

[14] Nymity. 2017. GDPR Compliance Toolkit. https://www.nymity.com/gdpr-toolkit.aspx

[15] H.J. Pandit and D Lewis. 2017. Modelling provenance for gdpr compliance using linked open data vocabularies. In *Proc of PrivOn*.

[16] Harshvardhan J. Pandit, Declan O'Sullivan, and Dave Lewis. 2018. Queryable Provenance Metadata For GDPR Compliance. *Proc. of SEMANTICS* 137 (2018), 262 – 268. http://www.sciencedirect.com/science/article/pii/S1877050918316314

[17] Tuan-Dat Trinh, Peb R Aryan, Ba-Lam Do, Fajar J Ekaputra, Elmar Kiesling, Andreas Rauber, Peter Wetz, and A Min Tjoa. 2017. Linked data processing provenance: towards transparent and reusable linked data integration. In *Proc. of the International Conference on Web Intelligence*. ACM, 88–96.

[18] Tuan-Dat Trinh, Ba-Lam Do, Peter Wetz, Amin Anjomshoaa, and A Min Tjoa. 2013. Linked widgets: An approach to exploit open government data. In *Proc. of Information Integration and Web-based Applications & Services*. ACM, 438.

[19] Tuan-Dat Trinh, Peter Wetz, Ba-Lam Do, Elmar Kiesling, and A Min Tjoa. 2015. Distributed mashups: A collaborative approach to data integration. *IJWIS* 11, 3 (2015), 370–396.

[20] Wil MP Van der Aalst. 2011. Process Mining. In *Process Mining*. Springer, 95–123.

On Preserving Sensitive Information of Multiple Aspect Trajectories In-House

Spyros Giotakis
University of Piraeus (UPRC)
Greece, sgiotakis@gmail.com

Nikos Pelekis
Department of Statistics and
Insurance Science
University of Piraeus (UPRC), Greece,
npelekis@unipi.gr

ABSTRACT

Nowadays, human trajectories are enriched with semantic information having multiple aspects, such as by using background geographic information, by user-provided data via location-based social media, as well as by data coming from various kind of sensing devices. This new type of multiple aspects representation of personal movements as sequences of places visited by a person during his/her movement poses even greater privacy violation threats. This paper provides the blueprint of a semantic-aware Moving Object Database (MOD) engine for privacy-aware sharing of such enriched mobility data and introduces an attack prevention mechanism where all potential privacy breaches that may occur when answering a query, are prevented through an auditing methodology. Towards enhancing the user-friendliness of our approach, we propose a mechanism whose objective is to modify the user queries that cannot be answered due to possible privacy violation, to 'similar' queries that can be answered without exposing sensitive information.

CCS CONCEPTS

• Information systems • Information systems applications • Spatial-temporal systems

KEYWORDS

privacy, anonymity, semantic trajectories, multiple aspect trajectories, query auditing

ACM Reference format:

Spyros Giotakis and Nikos Pelekis. 2019. On Preserving Sensitive Information of Multiple Aspect Trajectories In-House. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3317594

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317594

1 Introduction

Nowadays, the vast amount of spatiotemporal 'fingerprints' (i.e. trajectories) of individuals may prove to be a dangerous tool in the hands of a malicious user. So far, the scientific community has proposed various approaches to protect individual's privacy ([1] [2] [8] [10] [15]). Most of these studies focus on the spatiotemporal nature of this data, handling them as sequences of points on a geometric space, without considering that such raw data are usually enriched by additional information from the application context. However, the increasing need of analyzing mobility data has led to an advanced representation of trajectories enriched with contextual data from external data sources, thus transforming raw trajectories to the so-called semantic trajectories. A semantically-annotated trajectory, in short *semantic trajectory*, is considered as a sequence of stop episodes (i.e. places where the object remains "static") and move episodes (i.e. parts of the object's trajectory in between two stops) [11]. Each of them may contain additional annotations (i.g. home, cinema, work, etc.). Recently an even more advanced notion of *multiple aspects trajectories* [7] further enriches this kind of data with information coming from any kind of sensing devices (e.g. weather data, measurements from health apps, etc.).

This enriched representation of trajectories may pose even greater privacy violation threats. Consider for example a malevolent user who is able to detect places of interest (POIs), where a moving object has visited (e.g. hospital, betting office, etc.). This additional knowledge allows the inference of personal sensitive information of this specific individual. On the one hand, analyzing semantically-enriched movement traces of users can aid decision making in a wide spectrum of applications, but on the other hand, the disclosure of such data to untrusted parties may expose the privacy of the users, whose movement is recorded. Sharing user mobility data for analysis purposes should be done only after the data has been protected against potential privacy breaches.

The trend in the literature aims at protecting users' privacy by releasing an anonymized version of the original dataset ([1] [2] [6] [8] [10] [15]). These approaches assume that in the anonymized dataset a

malevolent user will not be able to link a specific user with a movement. In this paper, we employ a more conservative approach regarding privacy by assuming that data stay in-house to the hosting organization in order to prevent any privacy breach. We design a query-based auditing mechanism that can effectively identify and block a range of potential attacks that could lead to user identification or tracking, thus controlling the requested information that is released to third parties and ensure privacy-aware data sharing.

In more detail, in this proposal we make the following assumptions to tackle the problem. An enriched trajectory database contains, apart from the 'raw' trajectories, a number of episodes in which these trajectories were 'fragmented'. Every trajectory corresponds to one or more episodes. Each episode consists of data, describing its spatial range, its time range, the 'kind' of episode (i.e. Stop/Move) and a potential number of tags (annotation) which semantically enrich it. The four different pieces of information mentioned above are the data on which every user is able to apply several criteria through his/her query. Every query is a sequence of one or more independent sub-queries and each of them includes at least one of the four pieces of information previously mentioned as criteria. Every user receives as an answer to his query only the number of the trajectories that fulfill the criteria of his query and every such query as well as its answer is stored in the database for possible future reuse. It is considered that every annotation (tag) which may exist in the database is potentially known to the malevolent user. The proposed mechanism provides an answer only if k-anonymity principle is not violated w.r.t the user's current history.

Given the above, this paper makes the following contributions:

- We trace various types of attacks and thus privacy violations that malevolent users may try by querying the original enriched trajectory database.
- We device a query-based auditing mechanism that can effectively identify and block the potential attacks that could lead to user identification or tracking.
- We propose the *LENS* algorithm aiming at increasing user friendliness of the proposed mechanism by modifying the (original) query posed that cannot be answered due to privacy violation, to the 'nearest' query that can be safely answered.

The rest of the paper is structured as follows: Section 2 presents related work. Section 3 introduces different types of attacks of a malevolent user. Section 4 provides the auditing mechanism that handles the previously described attacks as well as the *LENS* algorithm. Finally, Section 5 concludes the paper.

2 Related Work

The *k*-anonymity principle [14] is the most common approach that has been adopted for the anonymization of both relational and mobility data. For mobility data, it states that a dataset must be anonymized so that every trajectory is indistinguishable from at least *k*-1 other trajectories.

Hoh and Gruteser [5] presented a data perturbation algorithm that is based on path crossing. Terrovitis and Mamoulis [15] consider datasets as sequences of places visited by users and proposed a suppression technique that eliminates the least number of places from a user's trajectory, so that the remaining trajectory is *k-ano*nymous. Abul et al. [1] proposed a *k*-anonymity approach that relies on the inherent uncertainty of moving objects whereabouts, where a trajectory is considered as a cylinder. The anonymity algorithm employs space translation and generates clusters of at least *k* trajectories. Each cluster of *k* trajectories forms an anonymity region and the co-clustered trajectories can be released. To achieve space-time translation, the authors proposed W4M [2], which uses a different distance measure that allows time-warping.

Nergiz et al. [10] proposed a coarsening strategy to generate a sanitized dataset that consists of *k*-anonymous sequences. Monreale et al. [8] proposed another anonymization approach that is based on the combination of spatial generalization and *k*-anonymity principle, while in [9] authors faced the problem of anonymizing semantic trajectories. To release a safe version of a semantic trajectory dataset, they propose a method that generalizes sequences of visited places based on a privacy place taxonomy.

On the other hand, in several sharing scenarios, many consider that data should stay in-house to the hosting organization and only a specific number of authorized users should have access on them. In this way, the hosting organization would be able to confront several possible legal restrictions, it could record the individuals who use the database and it would update the database whenever appropriate. Despite all the above, a mechanism is, in any case, needed in order to ensure that no sensitive information will be released during this process. Along this direction, methodologies have been proposed for disclosure control in statistical databases [3], but they support only count and/or sum queries.

In [4] the authors proposed an envisioned query engine where subscribed users had restricted access to the database to accomplish various analysis tasks and in [12] and [13] a privacy-aware query engine was introduced that can protect the trajectory database from potential attacks, while supporting popular queries for mobility data analysis. Both approaches deal with spatiotemporal trajectory databases not enriched, as in our case. This paper improves the approach in [6]. Finally, in [16] authors proposed a data stream management system, aiming at preserving users' privacy by enforcing Hippocratic principles.

3 Privacy Attacks

The main purpose of every attack of a malevolent user is to broaden his knowledge about an individual or a situation that interests him. This occurs when the attacker *raises his confidence* about an event that may be related to an individual, who is the 'target' or a situation for which he wishes to acquire more specific knowledge. Usually, a malevolent user has prior knowledge, i.e. time, place, type of episode and/or semantics (or any possible combination) about an individual.

3.1 Totally Overlapping Queries

Overlapping queries is a sequence of at least two queries posed by a user, having as a characteristic that the criteria of these successive queries are overlapping. We assume that either the queries have the same number of sub-queries and their criteria differ only in one dimension (space / time / semantics) or they differ in the number of sub-queries that each one contains keeping their 'common' sub-queries totally identical between them w.r.t. their criteria.

Spatial Overlapping. In this attack, the identity of a user can be revealed by posing overlapping queries, which differ in spatial dimension. The attacker poses a query and if the number of trajectories is at least k, he proceeds with one or more queries modifying each time only the spatial dimension, such that every time the new query contains the previous one. In fact, the most targeted way for a malevolent user to perform a 'successful' attack is to modify the spatial criterion of only one sub-query (or the same sub-query each time).

Let's assume that a user poses query Q_1 that contains n trajectories where $n \geq k$ and then Q_2 which returns as an answer $n+m$ trajectories, where $m<k$. The malevolent may conclude that the area corresponding to the difference between Q_1 and Q_2 contains m trajectories which is less than threshold k, thus privacy violation is occurred.

Consider the example depicted in Figure 1. A user poses query Q_1: *Find people starting from area A between [8.00-8.30am] and, then, stop at area B between [9.15-11.30pm].* This query contains two different sub-queries, each one able to provide an answer, if posed independently from the other. The same user poses query Q_2: *Find people starting from area A' between [8.00-8.30am] and, then, stop at area B between [9.15-11.30pm].* The answer of Q_1 contains 7 trajectories while the answer of Q_2 contains 8 trajectories. Thus, the malevolent can easily infer that only *one* person appears in the *area [A'-A] between [8.00-8.30am]* and subsequently stopped at *area B between [9.15-11.30pm].* By combining this knowledge with additional information, the malevolent can identify this person and extract the information that he was alone at that specific time and place.

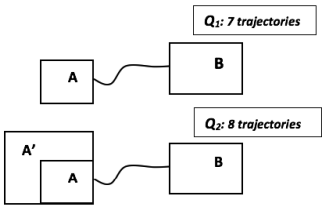


Figure 1: An example of two Overlapping Queries

The same reasoning, described above, can be followed regarding attacks applied by using two or more successive queries, which overlap in the temporal dimension.

Annotations Overlapping. In this case, a malevolent user may attempt to pose successive queries that take into account differences regarding only the semantic dimension. Let's examine the following example assuming that k=3. The user poses query Q_1 which consists of one sub-query and the answer corresponds to a specific spatiotemporal area that includes 7 *stop* episodes. During Q_2 the malevolent maintains the criteria of Q_1 but also adds *tag*='work'. The output of Q_2 contains 6 trajectories. The malevolent can conclude that only *one* entity was not working and he/she may be a workplace visitor.

Another example, using k=3, is this. Supposing that a malevolent user knows beforehand that a previously monitored/recorded person was found in a specific spatiotemporal area and had tag = 'fun' and this person claimed that he/she was accompanied by another (also monitored) person. The user poses query Q_1 regarding this area without adding any specific tag annotation and it returns 4 episodes. By posing Q_2 the user adds *tag*='work'. If the result of Q_2 contains 3 trajectories, then the recorded person is lying.

Different Number of SQs. In this attack, a malevolent user poses a sequence of queries which differ only in the number of the sub-queries that each one contains. Between two consecutive queries, assume that Q_1 contains n sub-queries and Q_2 contains $n+m$, where $m \geq 1$. In order to achieve an attack, the malevolent should know that his target 'participates' in the n sub-queries. Subsequently,

the malevolent may compare the number of the trajectories corresponding to the query consisting of the n sub-queries in relation to the one with the $n+m$ sub-queries. If the difference of the number of trajectories that constitute the answer of these two queries is less than k, this may result in revealing sensitive information about the target.

Consider the following example. The malevolent is aware of a target's home and working address and he/she intends to learn if the target returned at his residence after work. Assume that k=4. The user poses the query Q_1 consisting of two sub-queries that definitely 'contain' the target's trajectory (how many people stayed during the night in area A and during the following day were working in area B). The number of trajectories that fulfill Q_1 is 5. Q_2 contains exactly the same sub-queries with Q_1 along with a new sub-query that asks for those that returned after work back to area A during the night. If the answer of Q_2 returns 4 trajectories, then there is an 80% probability that the target spent indeed the night at home. This constitutes a significantly increased certainty compared to the 'safety limit' defined by the k-anonymity threshold. In other words, the malevolent should never be more than 25% certain about any situation 'recorded' in a database protected by a k=4 threshold. If the result again was 5, then he would be absolutely certain that the target returned home. We reasonably understand that if the user performed (only) the Q_2 from the beginning and regardless of the outcome, the result would be the same, but this *would not increase* his or her certainty about the behavior of the target in any way.

3.2 Multiple Intersecting Queries

This type of attack requires a sequence of, at least, 3 queries posed by a user. For simplicity reasons, let's assume that we are dealing with queries which comprise only one sub-query. The sub-query's criteria with which an attack can take place in this case are only space and time. We assume that the triple (at least) of the queries posed by the user differs either in space or in time each time, thus the malevolent user will be able to reach more specific and therefore alarming findings using the semantic database.

Consider the example depicted in Figure 2. Supposing that k=3, the user initially poses the query Q_1 whose spatial criterion corresponds to *area A* and it includes 5 episodes/trajectories. Then, he/she poses query Q_2 whose area corresponds to *area B* which covers part of *area A* along with some *extra area* and it returns 3 episodes. If a third query Q_3 (*area C*) corresponds exactly to the remaining part of *area A* that does not intersect with *area B* and it includes 3 episodes, then a privacy breach takes place. This is happening because, although none of the queries above do not overlap with each other, considering the fact that the 3 out of 5 episodes of *area A* are situated in *area C*, the user will be able to conclude that among the 3 episodes belonging to *area B* the two of them are situated within the *area A* and *only one episode* corresponds to the part of *area B* that does not intersect with *area A*.

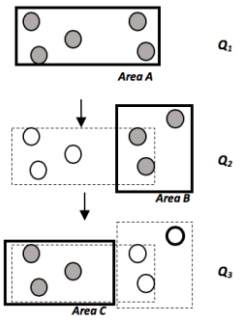


Figure 2: An example of three Intersecting Queries

Similarly, every time a user alters marginally only the temporal dimension between at least 3 successive intersecting queries and the spatial dimension remains the same, similar situations may arise.

4 Attack Prevention

To prevent the previously described attacks, an auditing mechanism is required to: (a) ensure that k-anonymity principle is not violated before answering each query; (b) protect certain episodes that may reveal sensitive information about entities and should be addressed in a specific way; (c) activate the *LENS* algorithm which allows the appropriate modification of the original query if k-anonymity principle is violated, in an attempt to make it acceptable; (d) allow the data owner to monitor and record constantly all the queries posed per user as well as the corresponding answers in order to compare them with every new query posed to the database.

4.1 Sensitive Episodes

Sensitive episodes correspond to known locations that contain particularly sensitive information and can possibly expose the identity of a user. We call these locations *sensitive* for a user as no such information should be disclosed to the attackers. In order to deal with user-defined sensitive episodes, the auditing mechanism initially does not count them as part of the query's answer set, but if the number of the non-sensitive episodes corresponds to at least k trajectories, then these sensitive episodes are allowed to be finally incorporated in the answer set.

4.2 LENS Algorithm

When a user poses a query to the database, he is willing to gain knowledge about whether there are semantic trajectories that

are answering the query w.r.t. certain criteria. If the number of the semantic trajectories, composing the result set, is less than the anonymity threshold *k*, the query is not safe to be answered. This ensures a first level of privacy protection.

The main idea of the proposed approach is that, instead of not providing an answer whenever *k*-anonymity principle is violated, an auditing mechanism should try to answer the query posed by a user in any case. In other words, the mechanism will provide an answer of the most 'similar' query to the original, that fulfills *k*-anonymity principle by relaxing conditions via generalization. Query relaxation enlarges the search range to include additional information. The output of this process is like a generalized query in one or more possible dimensions. Put differently, a user seeks to query an area, but the mechanism resolves it for a zoomed-out area that is generalized up to a permissible degree of analysis.

The main goal of implementing such an approach is to increase user friendliness and improve database functionality. A user can gain information without posing consecutive queries expanding repeatedly the criteria set until an answer is provided, because in this case there is an extra computational cost to the database and the possibility the user might not be given the best possible answer. To achieve this, the mechanism allows the generalization of one or more criteria, of the sub-queries that constitute the original query. The generalization may occur in the spatial, the temporal or even the semantic dimension.

Let's assume that the potential answer of an initial query *Q* consists of less than *k* trajectories. The output of the algorithm is a modified query *Q'*. If the modification process is successful, then the execution of the query *Q'* will result in at least *k* trajectories. The number of the sub-queries in both *Q* and *Q'* is of the same size. Each sub-query of *Q'* is either the same or generalized based on the corresponding sub-query contained in *Q*.

An obvious approach would be to apply the aforementioned method only in case that a query cannot be answered marginally w.r.t. to *k*-anonymity threshold. A threshold should then be defined. The mechanism, based on it, would be activated every time the query could not be answered. But in case the mechanism is activated when only few trajectories are missing from the answer set, a privacy breach may occur. A malevolent user can easily assume that the modified query does contain certain number of additional trajectories within the returned extra area. An obvious solution is to apply *LENS* algorithm, regardless of the number of trajectories needed to reach *k*-anonymity principle.

The goal of the algorithm is to modify one or more sub-queries, so that more episodes will be included per sub-query. To help the algorithm decide which episode is preferable to be included in the answer of the modified query, it should be able to compare the distortion that is caused on each sub-query, when trying to select between two or more candidate episodes. A unit, that calculates the distortion caused due to the generalization of one or more dimensions on each sub-query, is required. The distortion should be as low as possible to limit the generalization to the necessary level w.r.t. the initial query posed by the user.

The *LENS* algorithm takes as input a semantic trajectory database, the original query posed by a user that cannot be answered, the anonymity threshold *k*, the distortion limit value *dist* and a matrix *H*. The output of the algorithm is the modified query along with the corresponding sub-queries.

The algorithm after the initialization process (lines 1-2) continues with a loop phase where each sub-query of the original query is executed individually and the trajectories that comprise each one are retrieved (line 6). Each trajectory id of these trajectories is inserted into a matrix *(H)* along with the frequency indicating its appearance *(freq)* in all sub-queries (line 7). When a trajectory id first enters the matrix receive 1 as the value of *freq* and its frequency is increased by 1 every time the same trajectory (through its episodes) is found within the answer set of the sub-queries subsequently executed. Thus, *H* is a set of tuples each of which contains trajectory id (*traj_id*), frequency (*freq*) and the sub-queries (SQ_i). The maximum value that the counter *(freq)* can take in each record is equal to the number of the sub-queries. Consider as an example a query with three sub-queries. The counter for each trajectory in matrix *H* will receive a value ranging from 1 up to 3. If the counter receives the maximum value, it means that the episodes of this trajectory have been identified in all sub-queries, thus this trajectory can be returned as an answer to the overall modified query.

Based on matrix *H*, the algorithm detects the trajectory or trajectories with frequency (i.e., number of sub-queries), less than the maximum possible frequency, but at the same time with the highest value among the other trajectories in the matrix (line 9). Subsequently, a loop starts that ensures that, if no episode is found, the algorithm will search the trajectory that has the subsequent smaller frequency. This loop ends either when permissible episodes can be integrated, or if all the remaining trajectories from the matrix have been investigated and no episode is found (line 15). To define

the most appropriate candidate episodes, the algorithm employs a process called *Compute_Distortion_Units*. A metric function calculates the distortion caused in a sub-query in order to be modified and be able to include an episode from the trajectory. In case we have a distortion unit greater than a distortion limit (user-defined), the episode takes the tag INF and the algorithm proceeds with the next trajectory. Under these conditions the algorithm selects as preferable the episode that has the lowest distortion (line 13).

As a next step, the sub-query is modified in one or more dimensions to contain the episode that minimizes the distortion (line 17). The repetition ends (line 19) either if *k* trajectories have frequency equal to the number of sub-queries or if no episodes were integrated.

Algorithm 1: LENS

Input: (1) anonymity threshold *k*, (2) initial query with sub-queries Q = <SQ_1, SQ_2,..., SQ_n>, (3) a semantic trajectory database *D*, (4) distortion limit *dist*, (5) array H[tr_id, freq, SQ_1, SQ_2, ..., SQ_n]
Output: Q'= <SQ'_1, SQ'_2,..., SQ'_n>
Q' ← Q; H ← ∅
N_{tr} ← Count(Q')
repeat
    Something_Changed ← False;
    for i=1 **to** n **do**
        Execute_Query(in SQ'_i **out** tr_ids)
        Fill_Help_Table(in H, tr_ids **out** H)
    end for
    Find_Freq_Position(in H, n **out** i**)**
    episode_found ← False
    repeat
        Compute_Distortion_Units(in episode_found, H, i**)** **out**
        Select_best_candidate_episode(in H, dist, i **out** tr_id, ep_id**)**
        i ← i+1
    until episode_found **or** EOF
    if episode_found **then**
        Embed_New_Episode(in H, tr_id, ep_id, **in out** SQ'_i, Something_Changed)
        N_{tr} ← Count(Q')
until (not Something_Changed) **or** (N_{tr}=k)
Compute_Random_Number(in Rmin, Rmax **out** R)
Compute_New_Episodes(in Q', R **out** Q'**)**
return Q'

During the generalization process of the sub-queries, a privacy breach may occur. Let's assume that the spatial dimension of the area that the query covers is enlarged, so as to contain exactly *k* episodes. The spatial generalization should be the minimum possible in order to keep the distortion caused from this process as low as possible. To achieve this, most of the episodes that are added will appear in the borders of the modified area. The malevolent user, thus, will be more confident that, at least one episode exists, between the query posed and the modified query that is finally answered. In order to avoid such a violation, the modified query is expanded on each side by a randomly generated percentage *R* (line 21). Finally, we get as output the final modified query along with its sub-queries (line 22).

4.3 Query Auditing

The main goal of the *Query-Auditing* algorithm is to prevent the semantic database from answering a query to the user when it is possible that he may acquire further knowledge of a situation or a specific individual whose trajectories are stored on the database than he might have had before posing the query. The only way to avert this is to deny answering the query. In order to apply such a mechanism, it is necessary to keep the past user's queries along with the corresponding answers stored in the database.

Related works do not provide any answer when two queries posed by the same user are overlapping, aiming to prevent any privacy violation. In order to increase user friendliness and system functionality, we argue that the previous approach is very conservative and the algorithm should proceed to further examination before denying an answer. In other words, we find that it may be very strict to deny an answer without previously examining the extent that every query, ready to be answered, is overlapping to every query previously posed by the same user. More specifically, since each user has the right to receive an answer every time the k-anonymity principle is not violated, we argue that he still has the same right even if his current query overlaps a previous one, as long as the difference of these two queries' answers (i.e. the number of trajectories contained in each answer) equals or is greater than *k* threshold.

However, after applying this policy to the auditor, this may lead to a new type of potential privacy breach. More specifically, let's consider the following example assuming that a *k*=2 threshold is applied. Supposing there is a query Q_1 having only one sub-query and its temporal criterion defined between [8.00-11.00am] and it contains 5 trajectories. Next, if the same user poses a query Q_2, whose temporal dimension is [8.00-9.00am] and it contains 2 trajectories, the auditor should allow it to be answered, because the difference of these two answers is greater than *k*. Subsequently, if the same user tries a third query Q_3 whose temporal dimension is [9.00-10.00am] which contains also 2 trajectories, it seems that it should also be answered because although it overlaps (only) with Q_1, their difference is still greater than *k*. If so, a privacy violation will take place because this specific user will

eventually get to know that between [10.00-11.00am] only one entity has been recorded.

The solution we propose is to create dummy queries stored in the database along with the actual queries previously posed by the user and, in any case, the database will treat them as real queries posed in the past by him. Regarding the previous example, when the auditor replies to query Q_2, we create a dummy query with temporal criterion defined between [9.00-11.00am]. This would prohibit the user from querying Q_3 successfully. Then, we present specific examples, regarding the different types of overlapping queries that the auditor has to deal with.

Spatiotemporal Overlapping. Let's assume that k=3. A user poses initially the query Q_1: A→B→C (i.e. the query consists of 3 sub-queries) and subsequently the query Q_2: A→B→C'. The sub-queries C and C' are overlapping. The auditor examines the number of trajectories corresponding to the aforementioned queries and, let's assume, that Q_1 is satisfied by 7 semantic trajectories while the Q_2 by 4 ones. The difference of these two queries is equal to k, so the user will be allowed to receive answers for these two successive queries. However, in order to avoid future violation by the specific user, the auditor has to construct a fake query that will be stored in the database and corresponds to the spatiotemporal difference of the two queries Q_1 & Q_2 regarding their last sub-queries C & C', before releasing the final answer (i.e. Q_{fake}: A→B→[C∩C']). Thus, if the same user attempts a third actual query (4^{th}, including the fake one) Q_3: A→B→C'' and C'' totally overlaps [C∩C'], the same comparison between the number of trajectories of Q_{fake} & Q_3 will take place and if k-anonymity is still satisfied, then it will be answered too and a new fake query will be created, and so on. A similar approach to this also applies to multiple intersecting queries.

Annotations Overlapping. Our approach for a friendlier handling of queries in the case of overlapping tags is (see Sec. 3):

$$\|Q[\text{tag is null}]\| - \sum \|Qs[\text{tag is not null}]\| \geq k$$

Before answering the query, the auditor compares the number of trajectories corresponding to the query, without containing any specific tag, with the sum of the number of trajectories corresponding to all the other queries which contain tag(s) in their criteria and have already been answered. Obviously, all of the aforementioned queries are totally identical except only for the annotation criterion of only one sub-query and we note that the inequality also includes the query the auditor is examining at that time. If this inequality is true, then the auditor proceeds in answering the current query. There is, also, no need for generating a fake query when answering such queries.

Algorithm 2: Query-Auditing

Input: (1) anonymity threshold k, (2) initial query with sub-queries Q = <SQ_1, SQ_2,..., SQ_n>, (3) a semantic trajectory database D, (4) distortion limit dist, (5) user id u_{id}
Output: F_Q
$F_Q \leftarrow Q$; $Q_D \leftarrow \varnothing$
Remove_Sensitive_Episodes(in out F_Q)
if $\|F_Q\|$<k **then**
  LENS(in k, F_Q, D, dist, H **out** F_Q)
  if $\|F_Q\|$<k **then**
    return false
for each $Q_i \in$ D where user_id=u_{id} **do**
  for each $SQ_j \in Q_i$ **do**
    for each $SF_{Qm} \in F_Q$ **do**
      if SF_{Qm} **overlaps** SQ_j **then**
        Check_Suspicious_Query(in F_Q, Q_i **out** Is_Suspicious, Dummy_Query_Needed)
        if Is_Suspicious **then**
          if $\left|\|F_Q\| - \|Q_i\|\right| \geq$ k **then**
            if Dummy_Query_Needed **then**
              Create_dummy_query(in F_Q, Q_i **out** Q_D)
            else
              return false
    end for
  end for
end for
Add_Sensitive_Episodes(in out F_Q)
D ← D U Q_D
return FQ'

Different number of SQs. Let's assume that k=3 and a user poses the query Q_1: A→B→C→D and it is satisfied by 4 semantic trajectories. The same user poses Q_2: A→B→C and the answer contains 7 trajectories. Since the difference of these two queries that corresponds to the query A→B→C→not [D] is equal to k, no privacy violation can be caused. In fact, in this case there is no need for generating a fake query.

Algorithm's Description. The input of the algorithm is the anonymity threshold k, the initial query of the user along with the corresponding sub-queries, a semantic trajectory database D, the distortion limit value *dist* and the *id* of the user posing the query. The algorithm first executes query Q and gets the number of trajectories that make up the answer set (line 1). Then, the episodes that are considered as sensitive are defined and removed from it (line 2). If the number of trajectories is less than k, *LENS* algorithm, previously described, is called to modify the original query in an effort to provide an answer (lines 3, 4). If *LENS* algorithm cannot manage to modify the query w.r.t. a distortion threshold, no answer is provided to the user and the algorithm ends (lines 5, 6). On the contrary, if the original query Q or the modified query F_Q has equal or

more than k trajectories, the auditing mechanism continues to further investigate this query based on user's history.

The auditor proceeds by comparing every sub-query of F_Q with all the sub-queries that belong to queries posed from the user in the past. Every time two sub-queries are overlapping w.r.t their spatial, temporal or semantic dimension (including the case where these two are just intersecting each other) or they are totally identical, a procedure *Check_Suspicious_Query* is called to examine whether query F_Q is, in fact, a suspicious query compared to the current query Q_i which is a previously posed query stored in the Database (lines 10-11). If so, the auditor calculates the difference of the number of trajectories belonging to the corresponding queries (lines 12-13). If it is equal or greater than k (which means that the query will be answered in any case) and if the procedure *Check_Suspicious_Query* has previously determined the need of creating a dummy query, the auditor calls the procedure *Create_dummy_query* which creates the query Q_D (lines 14-15). Otherwise, the algorithm ends and no answer is provided to the user (line 17). Finally, if the query is to be executed, any sensitive episodes, that have been initially excluded, are added to the answer, the dummy query is stored in the database and the final answer is returned to the user (lines 21-23).

5 Conclusions

In this paper, we proposed an envisioned query engine able to provide safe answers to queries posed by users in semantic trajectory databases. Different types of privacy attacks have been addressed and an effective auditing mechanism able to prevent privacy breaches has been proposed. We subsequently introduce *LENS* algorithm which is able to modify an initially inacceptable query to the 'closest' one that can be safely answered, thus increasing the user friendliness of the engine. In the future, we plan to support multiple users access of the MOD.

6 ACKNOWLEDGMENTS

This work was partially supported by project MASTER (Marie Sklowdoska-Curie agreement N. 777695) that has received funding from the EU Horizon 2020 Program.

REFERENCES

[1] O. Abul, F. Bonchi, M. Nanni, 2008. Never walk alone: Uncertainty for anonymity in moving objects databases. In Proceedings of ICDE.

[2] O. Abul, F, Bonchi, M, Nanni, 2010. Anonymization of moving objects databases by clustering and perturbation. Information Systems, 35(8):884-910.

[3] N.R. Adam, J. C. Worthmann, 1989. Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4):515-556.

[4] A. Gkoulalas-Divanis, V.S. Verykios, 2008. A privacy–aware trajectory tracking query engine. SIGKDD Explorations, 10(1):40-49.

[5] B. Hoh, M. Gruteser, 2005. Protecting location privacy through path confusion. In Proceedings of SecureComm.

[6] D. Kopanaki, N. Pelekis, 2017. Privacy Preservation of Semantic Trajectory Databases using Query Auditing Techniques. In Proceedings of BIGQP @ EDBT/ICDT 2017).

[7] R. Mello, V. Bogorny, L. Alvares, Z. Santana, H. Luiz, C. Ferrero, A. Frozza, G. Schreiner, C. Renso. MASTER: A Multiple Aspect View on Trajectories. Transactions in GIS Journal, to appear.

[8] A. Monreale, G.L. Andrienko, N.V. Andrienko, F.,Giannotti, D. Pedreschi, S., Rinzivillo, S. Wrobel, 2010. Movement data anonymity through generalization. Transactions on Data Privacy, 3(2), pp. 91-121.

[9] A. Monreale, R. Trasarti, D. Pedreschi, C. Renso, V. Bogorny, 2011. C-safety: a framework for the anonymization of semantic trajectories. Transactions on Data Privacy, 4(2), pp.73-101.

[10] M. E. Nergiz, M. Atzori, Y. Saygin, 2008. Towards trajectory anonymization: a generalization-based approach. In Proceedings of Security and Privacy in GIS and LBS, SIGSPATIAL.

[11] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M.L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, 2013. Semantic trajectories modeling and analysis. ACM Computing Surveys (CSUR), 45(4), p.42.

[12] N. Pelekis, A. Gkoulalas-Divanis, M. Vodas, D. Kopanaki, Y. Theodoridis, 2011. Privacy-aware querying over sensitive trajectory data. In Proceedings CIKM.

[13] N. Pelekis, A. Gkoulalas-Divanis, M. Vodas, A. Plemenos, D. Kopanaki, Y. Theodoridis 2012. Private-HERMES: a benchmark framework for privacy-preserving mobility data querying and mining methods. In Proc. of EDBT.

[14] L. Sweeney, 2002. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 10(5):557-570.

[15] M. Terrovitis, N. Mamoulis, 2008. Privacy preservation in the publication of trajectories. In Proceedings of MDM.

[16] H. Wu, S. Xiang, W.S. Ng, W. Wu, M. Xue, 2014. HipStream: A Privacy-Preserving System for Managing Mobility Data Streams. In Proc. of MDM.Conference Name:ACM Woodstock conferenceConference Short Name:WOODSTOCK'18Conference Location:El Paso, Texas USAISBN:978-1-4503-0000-0/18/06Year:2018Date:June

Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape

Matthew Goerzen
Data & Society Research Institute
New York, NY, United States of America, matt@datasociety.net

Jeanna Matthews
Department of Computer Science
Clarkson University, Potsdam, NY, USA, jnm@clarkson.edu

ABSTRACT

In this paper, we analogize the practice of trolling to the practice of hacking. Just as hacking often involves the discovery and exploitation of vulnerabilities in a computer security landscape, trolling frequently involves the discovery and exploitation of vulnerabilities in a media or attention landscape to amplify messages and direct attention. Also like with hacking, we consider the possibility for a range of trolling personas: from black hat trolls who push an agenda that is clearly counter to the interests of the target, to gray hat trolls who exploit vulnerabilities to draw critical attention to unaddressed issues, and white hat trolls who could help proactively disclose vulnerabilities so that attack surface can be reduced. We discuss a variety of trolling techniques from dogpiling to sockpuppetry and also a range of possible interventions.

KEYWORDS

Trolling, hacking, media, ethics, politics, memes, journalism, media governance, social media

ACM Reference format:

Matthew Goerzen and Jeanna Matthews. 2019. Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317598

1 Introduction

In the early days of the Internet, users adopted the fishing term "trolling" to describe duplicitous postings designed to lure readers into engagement or reaction. While typically understood as a form of deceptive entertainment [1], the activity was also intelligible as a mode of governing other users, especially new users, in the Usenet discussion groups where the behavior was first documented [2]. A common pattern was for a veteran user to introduce a controversial topic that had been so extensively addressed in the past that only a new user would respond earnestly to the post. Other veteran users would know that the topic was a "troll," but new subscribers to the group would not--and thus might be baited into an extensive response, revealing their ignorance. This type of trolling could serve to identify "newbies" from group insiders and generally enforce community norms.

The concept of trolling has expanded to include a much wider range of behavior, from imageboard users raiding for "lulz" [3] [4], to hacktivists drawing media attention to activist causes [5], to state-sponsored propaganda and influence operations [6] [7] [8]. Trolling is often associated with online harassment, abuse, and disruption—especially with political or ideological intent [9] [10]. Trolls use controversy to sow chaos and amplify disinformation, bait journalists into reporting false information, and generally attempt to manipulate popular opinion [11]. But the label has also been applied to artists and online activists who exploit media dynamics to draw critical attention to the workings of socio-technical media systems [12] [13].

In this paper, we seek to frame trolling as the hacking of our shared media or attention landscape. We can view mass media and social media platforms as socio-technical systems to be secured. Through this lens, we argue that much like hacking often involves the discovery and exploitation of vulnerabilities in a computer security landscape, trolling typically involves the discovery and exploitation of vulnerabilities in our media or attention landscape. danah boyd has described activity of this kind as "hacking the attention economy" [14], while Rand Waltzman has called for "cognitive security" to address this mode of activity [15]. Most recently, Bruce Schneier and Henry Farrell have suggested expanding the scope of computer security to address "soft cyber" attacks such as politically-motivated trolling [16]. Attention hackers or trolls attempt to set the media agenda, directing journalists, social media users, and others to information that serves their interest. Often their techniques emulate or take advantage of platform dynamics designed to benefit advertisers and data-centric enterprises [17] [18]. The changing media landscape of mass media, social media, and algorithmic curation constantly introduces new vulnerabilities for attention hackers to exploit.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317598

2 Trolling Techniques

In Table 1, we list various trolling techniques. In most trolling techniques, a piece of disinformation is introduced as bait for a target or victim to ingest. The troll is typically looking for the target to take the bait and produce a reaction that will be observed by others. Often, the third-party observers or interpreters are meant to recognize the naivety of the target in accepting the bait as legitimate. The intended outcomes can vary from simply trying to embarrass the victim to provoking a particular political objective.

Table 1: Selected Trolling Techniques

Name	Description
Source Hacking / "Journobaiting"	Planting false claims or posing as sources to dupe mainstream media, especially in the wake of a crisis event.
Keyword Squatting	Associating a rarely used keyword or search term, especially one that becomes suddenly popular in the wake of a crisis event, with disinformation or propaganda.[a]
Denial of Service	Overwhelming a public space with information or data designed to promote exhaustion and disaffection, or generally frustrate sensemaking efforts.[b]
Sockpuppetry	The creation and management of social media accounts designed to appear as authentic individuals, typically using "data craft."[c]
Persona Management / Botherding	The co-ordination of multiple sockpuppet accounts or their algorithmic automation as a botnet.
Ironic Hedging / Bait and Switch	Using hateful or extreme language and imagery in a way that creates plausible deniability about intent and empowers messagers and some interpreters to downplay sincerity and seriousness.
Political Jujitsu	Soliciting attack from an opponent to elicit sympathy from political allies, ground victimization narratives, facilitate recruitment, and justify counterattack.
Controlling the Opposition	Using sockpuppet accounts to pose as representatives of an oppositional group.
Astroturfing	Using sockpuppet accounts to create the illusion of a bottom-up social movement or social outcry.
Wedge Driving	Inserting narratives designed to create divisive infighting among social groups. Often part of an overarching "divide and conquer" strategy.
Memetic Trojan Horses	The popularization of seemingly banal content that opens the Overton Window [d] by prompting commentary from mainstream journalists.
Deep Fakes [e]	Altering photographs and videos to change the original message, in a way that is difficult to detect.
Concern Trolling	Disingenuously expressing concern about an issue in order to derail discussion and damper consensus. Posing as a concerned ally or objective third party in order to make criticisms more palatable.
Brigading / Dogpiling	Bombarding a targeted individual or space with ad hominem attacks from multiple accounts.[f]
Conspiracy Seeding	Spreading "breadcrumbs" on social media and anonymous forums to nudge participants towards conspiracist thinking.
Algorithmic Gaming	Exploiting the functioning of an algorithm or related databases to elicit a result not intended by its designers. [g]

[a] This often involves SEO techniques, but can also involve the brute force appropriation of hashtags.

[b] The Grugq has described the release of a data trove of purportedly damaging information about Emmanuel Macron immediately before the 2017 French presidential election media blackout in these terms, as a "Data Denial of Service Attack" [19].

[c] "Data Craft" is the term used by Amelia Acker to describe the manipulation of metadata (geolocation, follower count, etc.) to enhance the authenticity of a fake social media account [20].

[d] The Overton Window refers to the range of political ideas available for address in public discourse.

[e] The term Deep Fakes comes from a subreddit devoted to producing videos using GAN techniques. It's come to stand in for a variety of image

manipulation techniques, including manual photoshopping—or what used to be called "shops."

[f] It should be noted that a number of individuals have pushed back on the idea that targeted harassment or abuse be described as "trolling," noting that the terms harassment and abuse already describe this behavior [21] [22]. However, we believe the prevalence of these techniques in subcultural trolling communities warrants their inclusion in this inventory.

[g] Of course, this raises questions about whether designer intent matters—i.e., whether an affordance is a "bug" or a "feature" and who gets to decide [23].

Trolls are often motivated to achieve specific political objectives. They may seek to draw attention to ignored issues ("agenda setting") or alternatively to waste attention and distract from issues. They may seek to pressure politicians, consumers, voters, businesses, or bureaucrats into action by shaping public attention, or they may attempt to draw a telling response as a way to increase transparency and accountability. They may wish to diminish an opponent's resources or morale, perhaps to exhaustion (e.g. by promoting "outrage fatigue" or dumping vast swathes of data, effectively "denying service"[1]). Some trolls seek to raise skepticism or to tear down all accepted sources of authority. Others simply want to cause trouble for the fun of it (lulz). Trolling can be done, as in the early Internet, to enforce community standards as a form of community building or governance, or to provoke deliberation regarding community norms.

Trolling can be surreptitious, but most trolling involves a revelatory moment where the deception is revealed or discovered, and the troll recognized. Common targets of trolling are journalists (and their readers), social media users (both allies and potential allies, and those perceived as opponents), companies, politicians, and regulators.

3 Vulnerabilities in Our Attention Landscape

In computer security, vulnerabilities are most often discussed in relation to technical systems like software code, hardware design, or network configurations and protocols.[2]

To consider where vulnerabilities lie in relation to trolling it is helpful to consider socio-technical systems. Here, vulnerabilities can present at multiple levels, typically where technical design overlaps with economic incentives, organizational factors, political objectives, and even psychological factors like cognitive bias. For our purposes, we will consider mostly social media, legacy news media, and the communities they service as the socio-technical systems particularly vulnerable to trolling.

In relation to news media, we can consider how a long-standing editorial predilection for novel, salacious, and spectacular content ("if it bleeds it leads") creates opportunities for trolls to bait journalists and set media agendas from below. We could also consider how editorial vulnerabilities have been exacerbated by the always-on, breaking news environment of continual online publishing and the competition for social media click-throughs. More complex to reckon with is the way many digital first media organizations themselves troll for user engagement through the production of "clickbait" headlines designed to circulate and capture attention on social media platforms.

Social media platforms have introduced new vulnerabilities due to their reliance on automated amplification and curation. For example, attention hackers can amplify their messages through the creation of bots that simulate the signs of organic human interest in a topic (e.g. liking, following, sharing, forwarding), or by reverse engineering algorithmic processes in order to game them. Understanding the metadata associated with messages and the ways in which online social media platforms use this metadata to decide which messages to amplify is key to identifying vulnerabilities that trolls can exploit [20]. In general, the anonymity available in many social media platforms offers both vulnerabilities and affordances for the amplification of messages.

Taking a broader scope, we could consider also that most forms of media or attention amplification systems harbor a critical vulnerability that exacerbates other vulnerabilities: the asymmetric cost to debunk false information. False, inflammatory, and hyper-partisan information may spread quicker than factual information, garner more attention, and be remembered longer [24] [25] [26]. Furthermore, attempts to debunk false information can unintentionally direct more attention and may counter-intuitively lead more people to believe the false information [27] [28].

Another example of a broad socio-technical vulnerability is "Poe's Law," a phenomenon named after Nathan Poe in 2005 when he observed that it is impossible to create a parody of extreme views so obviously exaggerated that it cannot be mistaken by some readers for a sincere expression of the parodied views [29]. Poe's Law can be exacerbated by design decisions (such as user anonymity) and more complex emergent factors (like "context collapse," where a publicly shared message intended for one audience is interpreted differently by another [30]).

In general, the economics of the ad tech economy and the ways in which algorithmic curation happens in online platforms is a source of many new vulnerabilities. For example, whenever a platform has not accumulated substantial history on a topic or search term, it is vulnerable to malicious actors dictating the output of the system [31]. Similarly, the economics of revenue for clicks can drive the creation and spread of divisive, false, and extreme content [32]. Nation-state actors have even created weaponized social media infrastructure for message amplification, that in many contexts swamps the organic signal of human discourse [33].

Considering all these examples, we can recognize that socio-technical vulnerabilities can appear in multiple places and at multiple scales. Some might consider that the entire existence of a platform like Facebook constitutes a vulnerability to the broader sphere of human interaction. Others might consider the

1 As Darin Barney has noted, excess information—even of an overtly political salience—can have a depoliticizing effect [45].

2 Of course, there are exceptions: organizational threat modeling often considers the impacts of social engineering attacks or the "human element," and many hacktivists advocate for a higher order conception of security aligned with human rights frameworks.

platform banal or even beneficial, but see the platforms policies around data enclosure and shareholder value as creating perverse incentives for its designers and advertisers. Others might consider that specific design features of Facebook—like the way its algorithms curate information—constitute vulnerabilities that ought to be fixed. Others still might suggest that only design features that do not work as intended constitute vulnerabilities. And still others might find fault with its content moderation policies. Determining what is a "bug" and what is a "feature" can often be subjective—and such judgment likely depends on an evaluator's propensity to benefit from the issue at hand.

4 Black Hat, Gray Hat, and White Hat Trolls

Within the computer security community, a clear distinction is made between black hat hackers who exploit computer systems to cause harm or generate profit for themselves and white hat hackers who exploit computer systems in order to proactively discover vulnerabilities that could be patched. We would like to propose a similar distinction for trolling.

A black hat troll could be understood to push a private agenda on their own that is counter to target's interests and counter to accepted standards of community governance. A white hat troll could be understood to identify, document, and reveal vulnerabilities selectively to platforms so as to reduce their attack surface and help align media systems and platforms with designer intent.

Black hat trolls would typically be using forms of engagement that are not prescribed by the designer of the system and white hat trolls would be looking for ways to disable forms of non-prescribed engagement through research and education.

Again, as with hackers, we further envision a category of gray hat trolls that use non-prescribed methods of engagement, but are motivated by a sense of public interest rather than an intention to harm or secure personal gain. They may view their actions as a way to demonstrate attack surfaces and thus increase pressure on gatekeepers to change their systems, or as a form of "hacktivism" that exploits the vulnerabilities of a system to further a perceived higher order goal of advocating for human rights or "human security."

White hat and gray hat trolls could perform penetration testing, engage in security research, and disclose vulnerabilities. They could also intervene by identifying bad actors, their incentives, and their methods. Furthermore, white hat and gray hat trolls might actively "troll back" or counter bad actors by "trolling the trolls."

The varied methods deployed by white hat hackers, black hat hackers, and gray hat hackers have inspired tremendous controversy in the computer security community—as debates around the full disclosure movement, hacktivism, bug bounty programs, and "hack back" policies can attest [34] [35] [36] [37], but these debates have also contributed to the establishment of functional computer security institutions. By schematizing trolling activities in this way, we could anticipate the generation of similarly productive debate regarding the ethics of trolling and the governance of socio-technical systems like social media platforms.

5 Amplification and the Design of Media Platforms

Defending against trolling requires consideration of acceptable forms of message amplification. For much of the history of 20th century pre-digital media, journalists and other mainstream media gatekeepers followed professional codes of ethics which typically included commitments to the accuracy of information, verification of sources, and the fair representation of multiple sides of an issue [38]. However, in practice these ethics are not always followed or enforced. And even when journalists operate within the bounds of these frameworks, they may be incentivized by thought of money, fame, power, political interest, or expertise to unduly amplify advantageous messages. Moreover, the same incentives can lead them to leave potentially important messages unamplified and under-represented.

Social media offered the promise of more democratized amplification of messages. However, in an information system in which anonymity can lead to multiple "votes" per person, we have encountered an emergent set of problems. Someone may have the right to free speech with their one mouth and body, but do they have the right to 20000 bots to amplify their message? Do they have the right to use sockpuppet to caricature those on the other side of an issue in a way that makes them appear ridiculous or extreme? When attention is scarce, does monopolization of attention by some constitute a denial of service attack on the rights of others' speech and an attack on the "attention sovereignty" of the limited attentional resources of individuals?

We have to grapple with the fundamental question of how we decide which messages should be amplified, and how best to gauge and represent "public opinion"—questions that have long concerned mass communications and journalism scholars, and only emerge anew in an era of "networked gatekeeping" [39]. Some have argued that the answer is educating individuals about fact checking and media literacy. However, is it really feasible for every individual to function as an investigative journalist, rigorously interrogating all claims that come into their media feeds? No one has the resources to verify/investigate every claim. When inflammatory lies are stickier and more profitable than real facts, it is difficult to imagine that fact-checking and individual action will be sufficient.[3]

If we believe that some form of gatekeeping collective attention will be necessary, then what kind? Human editors, algorithmic moderation? A preference for gatekeeping by automated systems guided by algorithms can lead to the overrepresentation of messages pushed out by those actors who have discovered ways to manipulate those systems in their favor. However, human gatekeepers, such as editors and content moderators, have demonstrated their own failings, blind spots, and incapacities. With the centralized human gatekeeping or

[3] Moreover, notions of "media literacy" may differ between different epistemic communities [40]

editing that was the norm in pre-digital mass media, ideas that emerged from marginalized communities and other disenfranchised groups often went completely unrepresented in mainstream mass media. We could consider the possibility of smaller communities selecting their own trusted gatekeepers. However, this kind of decentralized gatekeeping can lead to tribalism, where certain groups amplify fundamentally different messages from one another and lack the shared discursive objects that make co-operation and mutual understanding possible. We increasingly find ourselves in a world where different publics cannot even agree on the facts of our world, let alone align their opinions about those facts.

In the face of this fraught situation, a variety of possible interventions have nonetheless presented themselves. Below we identify a brief selection of possible top-down interventions and bottom-up interventions.

5.1 Top-down Interventions

There have been substantive discussions about the ways in which social media platforms like Twitter and Facebook could be improved (and possibly even regulated). Banning accounts that exhibit clear signs of algorithmic control (bots) is one way to reduce amplification through automation. However, this is only one source of manipulation. Disrupting economic incentives that reward click-through traffic on "fake news" or incentivize engagement for engagement's sake through regulation could be another important step [17]. Introducing additional hurdles for trust and verification of accounts (real-name policies, linking accounts to phone numbers or credit cards) is possible, but would also introduce a chilling of legitimate anonymous speech. Another category of top-down intervention could be considered "strategic silence," in which those individuals with editorial and editorial-like power in media and platform companies actively seek to avoid amplifying disinformation or even addressing it, to avoid generating more attention [27]. The user experience of a variety of platforms could benefit from the experimentation with different content moderation strategies [41]. Finally, enhanced transparency regarding moderation and amplification policies could promote good will among users, and reduce the opportunity space for political actors to claim victimization or unfair treatment. Of course, transparency on such matters could also enable those looking to brush up against the limits of acceptability as they manipulate or game platform dynamics.

5.2 Bottom-up Interventions

Not all interventions need to come from platforms or more traditional media gatekeepers. There are also a range of possible bottom-up interventions, including P2P moderation techniques like white lists and peer curated feeds. Users could introduce "countermemes" such as Godwin's Law to redirect online conversations [42], and promote active refusal to engage with disruptive influence ("don't feed the trolls") as a normative practice. Opting out of platform use is one possibility, but as many existing socio-technical media systems become increasingly integral to work and public life, both non-participation and early adoption of an emerging alternative could pose significant disadvantages to an individual [43] [44]. Users can collaborate to report content they see as negative for their community, effectively providing platforms with a rationale for removing content. Doxing those at the source of disinformation and shaming those that spread disinformation can also function as forms of community policing. Yet as stated previously, these techniques could further support the emergence of tribalism, and can themselves serve as modes of abuse or repressive gatekeeping.

6 Summary

In this messy modern world of message amplification, we see substantial benefit in understanding the history and techniques of trolling and how they serve to introduce and amplify messages. We see a role for white hat trolls to help platforms and media gatekeepers identify attacks and reduce their attack surface. We also see a role for gray hat trolls to point out vulnerabilities in the current media landscape, and to locate responsible parties and pressure them to change. If we want to see the benefits of the democratizing force of new media to amplify overlooked and under-represented voices, then we need to find ways to avoid the denial of service that comes from the manipulation of the processes and institutions responsible for gatekeeping and curating of our information environments.

7 ACKNOWLEDGMENTS

Thanks to Data & Society Research Institute for supporting this work.

REFERENCES

[1] J Donath, 1998 Identity and Deception in the Virtual Community, in Communities in Cyberspace. P Kollock and M Smith (Eds.). Routledge, London, England.
[2] M Tepper, 1996 Usenet Communities and the Cultural Politics of Information, in Internet Culture. David Porter (Ed.). Routledge, New York, NY, USA.
[3] G Coleman, 2012 Phreaks, Hackers, and Trolls: The Politics of Transgression and Spectacle, in The Social Media Reader. New York University Press, New York, New York, USA.
[4] W Phillips, 2015 This Is Why We Can't Have Nice Things. MIT Press, Cambridge, Massachussets, USA.
[5] G Coleman, 2014 Hacker, Hoaxer, Whistleblower, Spy. Verso, London, England.
[6] SC Woolley and PN Howard (Eds.). 2018. Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media (Oxford Studies in Digital Politics). Oxford University Press, Oxford, England.
[7] JC Ong and JVA Cabañes, 2018 Architects of Networked Disinformation. University of Leeds.
[8] C Paul and M Matthews, 2016 The Russian "Firehose of Falsehood" Propaganda Model. Rand Corporation.
[9] S Herring, K Job-Sluder, R Scheckler, and S Barab. 2002 Searching for Safety Online: Managing "Trolling" in a Feminist Forum. The Information Society, 18(5), 371-384.
[10] K Mantilla, 2015 Gendertrolling: How Misogyny Went Viral. Praeger, Santa Barbara, California, USA.
[11] A Marwick and R Lewis, 2017 Media Manipulation and Disinformation Online. Data & Society Research Institute.
[12] L Knuttila, 2016 Trolling Aesthetics: The Lulz as Creative Practice (Dissertation). York University, Toronto, Ontario, Canada.
[13] M Goerzen, 2016 Critical Trolling (MA Thesis). McGill University, Montreal, Québec, Canada.
[14] d boyd. 2017 Hacking the Attention Economy. Points blog, Data & Society Research Institute.
[15] R Waltzman. 2017 The Weaponization of Information. US Senate Testimony, Rand Corporation.
[16] B Schneier and H Farrell, Common Knowledge Attacks on Democracy. Berkman Klein Center Research Publication No. 2018-7.

[17] A Nadler, M Crain, and J Donovan. 2018 Weaponizing the Digital Influence Machine. Data & Society Research Institute.
[18] JA Braun and JL Eklund. 2019 Fake News, Real money: Ad Tech Platforms, Profit-driven Hoaxes, and the Business of Journalism. Digital Journalism, 7(1), 1–21.
[19] T Grugq. 2017 A Last Minute Influence Op by Data DDOS. Medium.
[20] A Acker. 2018 Data Craft: the Manipulation of Social Media Metadata. Data & Society Research Institute.
[21] J Bartlett. 2014 OG Internet Trolls are Upset their Hobby's Been Ruined. Vice.
[22] W Phillips. 2016 Donald Trump is Not a Troll. Slate.
[23] M Wolf, K Miller, and F Grodzinsky. 2017 Why we should have seen that coming: comments on Microsoft's tay 'experiment,' and wider implications. SIGCAS Comput. Soc. ACM SIGCAS Computers and Society, 47(3), 54–64.
[24] S Vosoughi, D. Roy, and S. Aral. 2018 The spread of true and false news online. Science, 359(6380), 1146-1151.
[25] M Sally Chan, CR Jones, KH Jamieson, and D Albarracín. 2017 Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. Psychological Science, 28(11), 1531–1546.
[26] A Friggeri, L Adamic, D Eckles, and J Cheng. 2014 Rumor Cascades. In Eighth International AAAI Conference on Weblogs and Social Media.
[27] W Phillips. 2018 The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators. Data & Society Research Institute.
[28] J Donovan and d boyd. 2018 The Case for Quarantining Extremist Ideas. The Guardian.
[29] SF Aikin. 2013 Poe's Law, Group Polarization, and Argumentative Failure in Religious and Political Discourse. Social Semiotics, 23(3), 301-317.
[30] AE Marwick and d boyd. 2011 I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. New Media & Society, 13(1), 114-133.
[31] M Golebiewski and d boyd. 2018 Data Voids: Where Missing Data Can Easily Be Exploited. Data & Society Research Institute.
[32] A Smith, and V Banic. 2016 Fake News: How a Partying Macedonian Teen Earns Thousands Publishing Lies. NBC News.
[33] PN Howard, B Ganesh, D Liotsiou, J Kelly, and C François. "The IRA and Political Polarization in the United States, 2012-2018." Computational Propaganda Research Project.
[34] A Arora, R Krishnan, A Nandkumar, R Telang, and Y Yang. 2004 Impact of Vulnerability Disclosure and Patch Availability-An Empirical Analysis. Third Workshop on the Economics of Information Security 24, 1268-1287.
[35] S Vegh. 2005 The Media's Portrayal of Hacking, Hackers, and Hacktivism Before and After September 11. First Monday, 10(2).
[36] T Ring. 2015 White Hats Versus Vendors: The Fight Goes On. Computer Fraud & Security, 2015(10), 12-17.
[37] V Jayaswal, W Yurcik, and D. Doss. 2002 Internet Hack Back: Counter Attacks as Self-Defense or Vigilantism?. IEEE 2002 International Symposium on Technology and Society (ISTAS'02). Social Implications of Information and Communication Technology, 380-386.
[38] Society of Professional Journalists. 2014 Code of Ethics.
[39] K Barzilai-Nahon. 2008 Toward a Theory Of Network Gatekeeping: A Framework For Exploring Information Control. Journal of the American Society for Information Science and Technology, 59(9), 1493-1512.
[40] F Tripodi. 2018 Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices. Data & Society Research Institute.
[41] R Caplan. 2018 Content or Context Moderation? Data & Society Research Institute.
[42] M Godwin. 1994 Meme, Counter-meme. Wired, 2(10), 10.
[43] M Vigil-Hayes, J Matthews, A Acker, and D Carter. 2018 Reflections on Alternative Internet Models and How They Inform More Mindful Connectivity. ITU Journal, ICT Discoveries, Special Issue 2, Data for Good.
[44] N Casemajor, S Couture, M Delfin, M Goerzen, and A Delfanti, 2015 Non-participation in Digital Media: Toward a Framework of Mediated Political Action. Media, Culture & Society, 37(6), 850-866.
[45] D Barney. 2014 Publics without Politics: Surplus Publicity as Depoliticization. In Publicity and the Canadian State, edited by K Kozolanka. University of Toronto Press, Toronto, Ontario, Canada.

Trust and Trustworthiness in Social Recommender Systems

Taha Hassan
Computer Science Department, Virginia Tech
Blacksburg, VA
taha@vt.edu

D. Scott McCrickard
Computer Science Department, Virginia Tech
Blacksburg, VA
mccricks@vt.edu

ABSTRACT

The prevalence of misinformation on online social media has tangible empirical connections to increasing political polarization and partisan antipathy in the United States. Ranking algorithms for social recommendation often encode broad assumptions about network structure (like homophily) and group cognition (like, social action is largely imitative). Assumptions like these can be naïve and exclusionary in the era of fake news and ideological uniformity towards the political poles. We examine these assumptions with aid from the user-centric framework of *trustworthiness* in social recommendation. The constituent dimensions of trustworthiness (diversity, transparency, explainability, disruption) highlight new opportunities for discouraging dogmatization and building decision-aware, transparent news recommender systems.

KEYWORDS

Recommender systems, social recommendation, collaborative filtering, trust, transparency, explainability, diversification, quality

ACM Reference Format:
Taha Hassan and D. Scott McCrickard. 2019. Trust and Trustworthiness in Social Recommender Systems. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3317596

1 INTRODUCTION

In the 2016 US Presidential Election, social media was one of the key venues for dissemination of fake news, with aggregate viewership of fake news strongly correlated with aggregate voting patterns [3]. These patterns hint at deeper geographical and sociopolitical groupings in the US (see figure 1). Differences in media habits by political affiliation have also been the subject of attention of late. A Pew Research Center study [1] found little evidence of overlap between news sources frequented and trusted by self-identified consistent liberals and conservatives. Conservatives gravitated towards a small number of news outlets, often just one, and exhibited distrust of a majority of the mainstream news sources. Liberals consulted a wider variety of news outlets, with social issue-based choices. In addition, consistent liberals were more likely to 'unfriend' or block someone because of their political views, while consistent conservatives were more likely to have friends who agree with their political views. According to the Pew Center study, Facebook draws twice as many consumers of news on politics and government relative to Yahoo or Google News, and about 40% of the US population would seek these news stories on Facebook week before the survey. A number of social recommender systems serving content as varied as pages, groups, events and jobs, work behind-the-scenes at Facebook. User interaction with a social recommender encodes interesting technical, behavioral and demographic trends, especially in the era of increasing partisan rifts. This study explores key limitations of

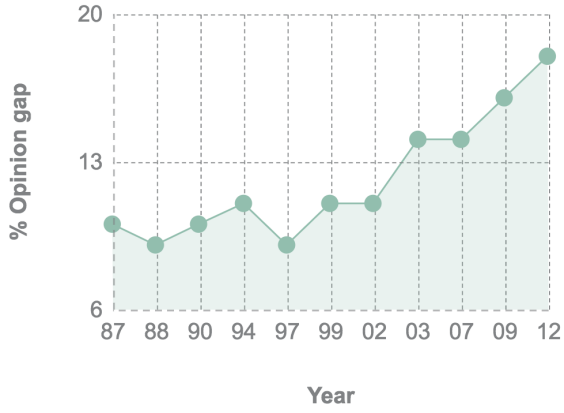


Figure 1: Partisan differences in political values (average percentage difference between Republicans and Democrats) on a 40-point values questionnaire [1]

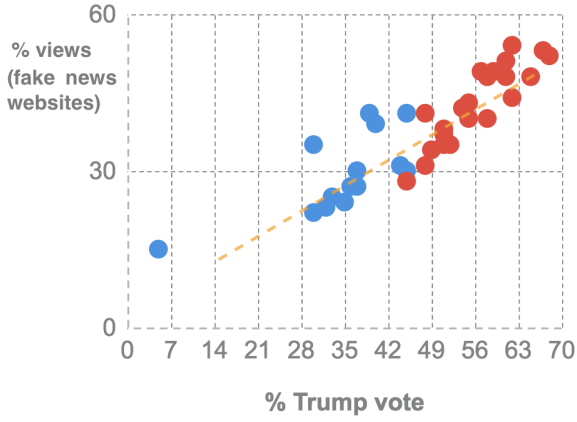


Figure 2: Voting behavior (by state) and average daily viewership of visitors to websites serving fake news. Blue represents a state-wide Democratic victory, red represents a Republican victory [3]

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317596

social recommendation from a critical-theoretic standpoint and provides new design considerations for their future.

The next sections review the conventional notion of trust-awareness, present a critique of this notion within the context of political polarization and explore issues of transparency, explainability and diversity for social news recommenders.

2 RELATED WORK

Implementations of recommender systems have been traditionally attributed to one of two approaches: the model-driven and the data-driven [8]. The first, known as collaborative filtering (CF) refers to modeling the users and items jointly, by a shared space of latent feature vectors [4]. The second, neighborhood modeling (NM), identifies user-user and item-item similarities directly. In practice, these two approaches have similar ancestry and can be put together in hybrid approaches.

Trust-based recommender systems are concerned with learning the preferences of trustworthy social neighbors, or 'friends' of an individual user, as well as the mistrusted 'foes' [12]. These preferences inform the latent features inferred in CF, such that features for an individual are ranked closer to his or her friends' features, rather than the foes. In [9], learning-to-rank models minimize a loss function on a personalized ranking function. Using trust and mistrust relationships proves effective in combating the sparsity inherent in users' preferences data.

User-perceived quality of recommender systems' output is a broad way to evaluate anything from aggregate emotional impact to perceived relevance and variety. User-perceived variety or diversity in the recommender output is a related, important notion. Authors in [5] explore an organization interface (ORG) as opposed to a list interface for a top-N style recommender. This interface clusters and annotates subgroups of the recommender output. These annotations go beyond category labels and express tradeoffs in product quality and price. ORG ranked better than lists in perceived ease of use and diversity. Studies like [15], however, note that statistical accuracy of such recommendations might be lower, even as they are rated better in quality by field trial participants.

3 TRUSTWORTHINESS

3.1 A Critique of Trust

User feedback in recommenders, be it explicit (ratings, reviews) or implicit (click behavior, browsing habits) is often sparse and non-generalizable. This sparsity has driven the need for the conventional notion of trust. Often leveraged in CF, this notion of trust assumes homophily in the social network, that like connects with or favors like [3]. Trust-aware CF, therefore, ranks the judgment of trustworthy friends as consistently higher than those of untrustworthy peers and foes. From a broad critical-theoretic standpoint, this construction of trust can be arbitrary and cynical. Note that sparsity encodes an emergent market need for showing relevant news stories rather than in-depth or disruptive ones. Also note that, the qualifier 'emergent' helps avoid the trendy critique of reverse-cynicism. It refers to the aggregate of choices by a large number of people potentially exercising their free will, choices such as reviewing or rating items less frequently than idly browsing a news feed.

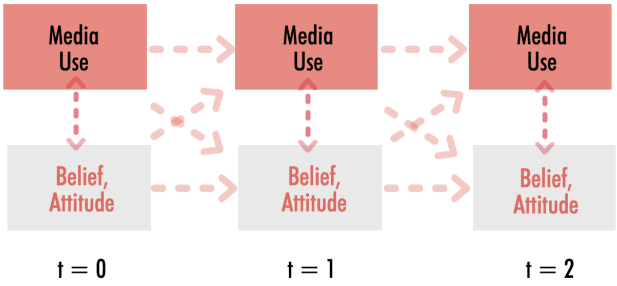


Figure 3: A three-way, lagged model of behavior reinforcement spirals [13].

To expand this critique, we can turn to a large volume of literature in the behavioral sciences domain [14][13] that examines the degree and scope of recommendation influence on online choices. Recall how consistent conservatives and liberals have strong selective tendencies in choosing whom to be friends with and what media sources to trust. The author in [13] presents a theoretical framework for bias reinforcement in social media (see figure 3). A closed feedback loop aims to maximize the use of a single social medium in the absence of competing interests. While a statistical analysis of these models is beyond the scope of the study, there are important conceptual aspects to consider here. One, consider the multitude of paths that can lead to feedback loops of gratification. Two, consider the assumptions behind nominal categories like media use, such as that of a closed system. In practice, institutions like marriage, family and religion, as well as competing media sources might help limit the scope of this reinforcement.

While there is some prior work on longitudinal trends in recommender output and rating diversity [10], reinforcement loops have not been formally studied from a recommender UX perspective (to the best of the author's knowledge). It is helpful, therefore, to reflect on the interactions that facilitate such loops. Authors in [7] describe a user-centric framework for perceived effort, accuracy and satisfaction of a recommender systems user (see figure 4). A certain dual causation exists between *experience* and *interaction* in this UX model. *Experience* can inform and modify *interaction*, as much as *interaction* informs and modifies *experience*. We can thus, leverage the spiral model in [13] to conceptualize bias-reinforcement as an undesirable byproduct of this duality. Furthermore, the study reports that the users who rate the recommendations as 'effective' have on average, a lower total viewing time and total number of watched clips. They do not necessarily consume a lesser volume of information, but less of their time is spent browsing than it is on watching recommended clips.

Another important distinction is the so-called 'greedy' approach to search. Some users are looking to find the 'best' possible recommendation while browsing, while others are looking to find the first perceived optimal recommendation. Domain expertise and trust affect these decisions [6]. In that, the tendency to favor sophisticated control (like, user-defined weights and order) over sort-only control correlates well with domain expertise. An 'average' user is likely a lot more passive in taking a recommender system's output at face value. Authors in [10] report that for long-term users of the

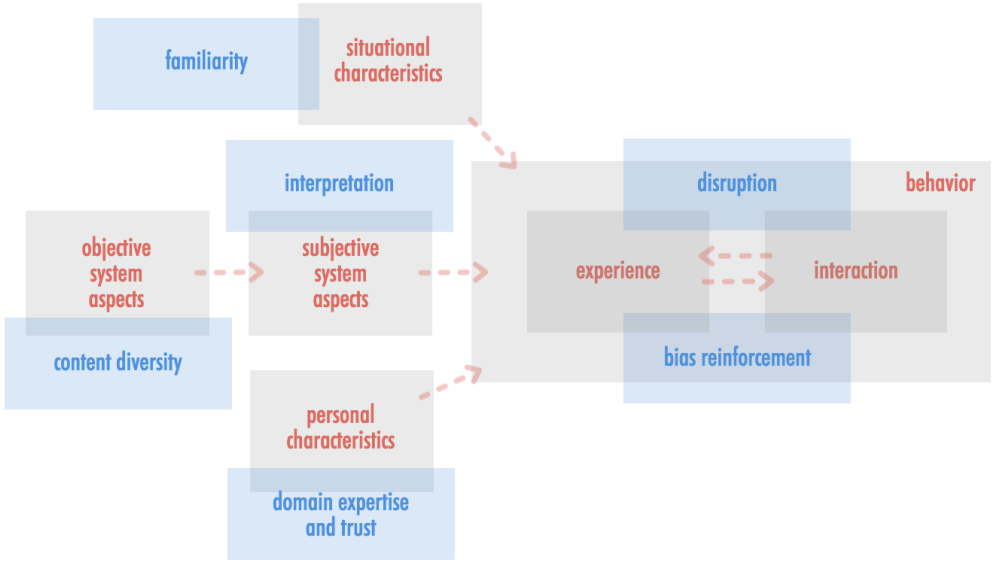


Figure 4: A recommender systems UX framework [7] from the perspective of trust and disruption.

MovieLens recommender, the diversity of recommender output and of explicit user ratings narrows over time. However, the drop in diversity is greater for users who tend to ignore the recommender output whilst rating content relative to those who do not.

This calls for design considerations that go beyond the local-neighborhood approach to recommendations. The output of a recommender system should help disrupt rather than reinforce feedback loops [2]. This involves a move towards explainability and transparency, as well as diversification and provocation.

3.2 Conceptual Dimensions of Trustworthiness

In the broadest sense, a trustworthy recommender is a recommender that enables transparent and interpretable interaction against the rising current of dogmatization and partisan antipathy. The key conceptual dimensions are as follows.

3.2.1 Diversification. Diversification for a news recommender improves the potential for a given user's encounter with alternative perspectives on a given social issue and disrupts the feedback loops explored in the previous section. There is room for creating a notion of diversity using a quality ranking for each news article. Google News, for instance, features 'in-depth' articles for a subset of search queries based on automated strategies. Similarly, Amazon.com recommenders arrange reviews by usefulness based in part on explicit feedback by readers of said reviews. A diversified recommender should recognize the degree of substantiation, i.e., it should be able to distinguish between an 'in-depth' analysis or news story versus a 'broad overview' and present a mix of labeled samples from each. Empirical studies in [6] and [7] associate the tendency to tag with a user's trust of the recommendation system and lack of privacy concerns. Therefore, manual tagging of article quality needs to be evaluated for agreement between independent labelers and between labelers and third-party fact-checking agencies.

3.2.2 Disruption. Model-level disruption is a more intrusive strategy that can potentially challenge the neighborhood assumption of recommender systems and create more provocative collections of news and analyses. The present state of recommender systems makes it difficult to get recommendations on a specified social perspective. In other words, we cannot ask a recommender system to fetch say an anarcho-syndicalist or a second-wave feminist perspective on a social issue or a news story, unless there is explicit or implicit history of a user browsing articles similar to such perspectives. However, consider the notion of transfer learning. Transfer learning for CF [11] allows cross-domain application of machine-learnt models and sparsity reduction. There is, however, no transfer-learning framework that uses a user-specified target domain to rank and retrieve recommendations, to the best of the author's knowledge.

3.2.3 Explainability and Interpretation. Another facet of explainability (as well as transparency) is decision-awareness for the user. Netflix's movie recommender, for instance, includes annotations on why it ranked a given movie or a TV show highly among the recommended content. These include references to actors, directors, genres and relevant decades, among others. A news recommender, especially one concerned with popular conversations, can potentially increase the users' trust and confidence in its output using explanatory annotations. This might especially be true with automated flagging of ad spam, content involving bullying and online predatory behavior, or NSFW images and video data.

4 FUTURE WORK

This study explores the theoretical foundations of trust-aware ranking in social recommenders. A key frontier in building such systems is verifying the user-perceived quality of disruptive interaction. Therefore, crowd-sourced tagging of article quality and authenticity, third-party verification, progressive disclosure of perspectives in a conversation and disruptive personas are subject of future empirical investigations by the author.

5 CONCLUSION

Disruption in recommender systems can take one of many forms. However, the relatively new interest in this domain implies a lack of frameworks, empirical studies, field trials and ethnographic work. Culture-awareness is an ambitious frontier for recommender systems of the future, and one that encompasses ideas like model differentiation, language preservation and community building. Diversification, transparency and disruption are key building blocks for tools that enable these ideas.

REFERENCES

[1] Jocelyn Kiley Amy Mitchell, Jeffrey Gottfried and Katrina Eva Matsa. 2014. Pew Research Center: Political polarization and media habits. Retrieved 2017 from http://www.journalism.org/2014/10/21/political-polarization-media-habits/

[2] Shaowen Bardzell, Jeffrey Bardzell, Jodi Forlizzi, John Zimmerman, and John Antanitis. 2012. Critical design and critical theory: the challenge of designing for provocation. In *Proceedings of the Designing Interactive Systems Conference*. ACM, 288–297.

[3] Adam Fourney, Miklos Z Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2071–2074.

[4] Taha Hasan, Naveed Arshad, Erik Dahlquist, and Scott McCrickard. 2017. Collaborative filtering for household load prediction given contextual information. In *Proceedings of the 2017 SIAM Workshop on Machine Learning for Recommender Systems (MLRec)*.

[5] Rong Hu and Pearl Pu. 2011. Enhancing recommendation diversity with organization interfaces. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 347–350.

[6] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 141–148.

[7] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[8] Yehuda Koren and Robert Bell. 2015. Advances in collaborative filtering. In *Recommender systems handbook*. Springer, 77–118.

[9] Paolo Massa and Paolo Avesani. 2007. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*. ACM, 17–24.

[10] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. ACM, 677–686.

[11] Weike Pan, Evan Wei Xiang, Nathan Nan Liu, and Qiang Yang. 2010. Transfer Learning in Collaborative Filtering for Sparsity Reduction.. In *AAAI*, Vol. 10. 230–235.

[12] Dimitrios Rafailidis and Fabio Crestani. 2017. Learning to Rank with Trust and Distrust in Recommender Systems. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 5–13.

[13] Michael D Slater. 2007. Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication theory* 17, 3 (2007), 281–303.

[14] Haiyi Zhu and Bernardo A Huberman. 2014. To switch or not to switch: understanding social influence in online choices. *American Behavioral Scientist* 58, 10 (2014), 1329–1344.

[15] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 22–32.

Empirical Analysis of Bias in Voice-based Personal Assistants

Lanna Lima*
lima.lanna@edu.unifor.br
lannalima.br@gmail.com
Universidade de Fortaleza
Fortaleza, Ceará

Vasco Furtado
Universidade de Fortaleza
Fortaleza, Brazil
vasco@unifor.br

Elizabeth Furtado
Universidade de Fortaleza
elizabeth@unifor.br

Virgilio Almeida
Computer Science Department, UFMG
virgilio@dcc.ufmg.br

ABSTRACT

Voice-based assistants are becoming increasingly widespread all over the world. However, the performance of these assistants in the interaction with users of languages and accents of developing countries is not clear yet.

Eventual bias against specific language or accent of different groups of people in developing countries is maybe a factor to increase the digital gap in these countries.

Our research aims at analysing the presence of bias in the interaction via audio. We carried out experiments to verify the quality of the recognition of phrases spoken by different groups of people. We evaluated the behaviour of Google Assistant and Siri for groups of people formed according to gender and regions that have different accents. Preliminary results indicate that accent and mispronunciation due to regional differences are not being properly considered by the assistants we have analyzed.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Laboratory experiments**; **Empirical studies in HCI**.

KEYWORDS

voice-based personal assistants, bias, machine learning.

ACM Reference Format:
Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical Analysis of Bias in Voice-based Personal Assistants. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317597

1 INTRODUCTION

Audio interaction will be the fastest and most convenient way for many purposes compared to using tactile monitors or remote controls [6]. Voice-basec assistants are rapidly becoming widespread and is reaching mainstream audiences in US and countries in Europe and Asia [8].

Voice-based wearables, audio, and video entertainment systems are examples of that. Also, voice control systems will soon be the norm in vehicles. Google Home and Amazon's Alexa are making popular the idea of a "`smart home`" across millions of households in the US. ComScore[8] predicts that by 2020, half of all our searches will be performed by voice. This scenario will be impactating in countries like Brazil, where a large part of the population is functionally illiterate [4]. The high adherence of Brazilians to Whatsapp (120 million active users/ month) using communication via audio is indicative of this [3].

Despite this enlightening perspective, voice-recognition technology has imperfections. The Wired Magazine recently exemplified how accent and gender differences might difficult human interaction [6]. A typical database of American voices, for example, would lack poor, uneducated, rural, non-white, non-native English voices. These kind of problems occurs in US with its large immigrant population, but is particularly critical in the developing world.

In Brazil and India, for instance, countries with large territories, geolinguistic variations are strongly related to the socioeconomic classes of the population. In Brazil, the southern region is known as the industrial center of the country, while the northern region possesses a less wealthy population. The economic issues influence the access to technology as well as the importance of the linguistic variations, that can provoke bias in the audio interaction.

Our research aims to analyze the presence of bias in the interaction via audio. The first step towards this was through an experiment to verify the quality of the recognition of phrases spoken in Portuguese by people of different genres and originating from different regions with different accents. Google Assistant and Apple's Siri were the assistants chosen.

Two questions have driven the research:

1: Does user's gender affect the understanding of the interaction system via audio?

2: Does the user's accent affect the understanding of the interaction system via audio?

An experiment was conducted in a Lab of a brazilian University with 20 volunteers, with the objective of elucidating these research questions. Preliminary results indicate that regional accent issues are not being properly considered by Google and especially by Siri.

2 RELATED WORK

Bias from audio interaction is a theme relatively few explored in the literature. It cannot be strictly classified as a bias of interaction

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-9999-9/18/06.
https://doi.org/10.1145/3308560.3317597

cf. [1], because the latter typically concentrate on the user interface and the user's own self-selected, biased interaction. Bias in the interaction is typically studied by analysing the user actions (like mouse movement and number of clicks). As for the audio interaction, this kind of technique is useless. The bias in the audio interaction comes from the process of training the assistant. It is a kind of data bias [1] originated from the creation of unbalanced databases that exclude parts of a population. Also, the interaction via audio depends rather on the user context and features than in their action and choices. Different speaking styles of human beings (i.e. the language accent, ethnic origin, emotion, gender, age) and the user's physical location are example of this.

Yanli Zheng and colleagues [12] worked on speech recognition for accents on Chinese. They evaluated the accuracy of Automatic Speech Recognition (ASR) of Interactive Voice Response systems (IVR), and proposed strategies to improve the accuracy of these systems. Similarly, reference [11] studied the difference in regional accent of English in England, America, Australia and India. A systematic view of ASR systems was done in [9] but the features used in the comparison of the systems are focused on their performance such as speed in real-time and response time. The study is a useful source for developers of devices like thermostats, doorbells, light bulbs and car accessories who want to add IVR (like virtual assistants) to the devices.

On the other hand, reference [10] evaluated the accuracy of speech recognition of a talker by a human (the listener). They show that listeners identify voices more accurately in their native language than in an unknown, foreign language. The talker identification [7]is studied, varying aspects related to the listener competence in the language to be spoken (such as her/his language familiarity, age, early language experience, cognitive skill, etc.).

Bellegarda [2] investigated the model behind Siri and its ability to infer knowledge. He concluded that ontology-based systems, such as Siri, are better suited for initial deployment in well-defined domains across multiple languages, but must be carefully tuned for optimal performance. Data-driven systems have the potential to be more robust, as long as they are trained on enough quality data.

Despite this related work, none in the literature has focus on bias due to regional accents within the country and the relation that this can imply in economic differences. Our study in this article with an experiment with participants speaking in Brazilian Portuguese for popular assistants is a step towards this.

3 EMPIRICAL EVALUATION

3.1 Methodology

An experiment was designed to verify the understanding of voice assistants in smartphones. The understanding was measured from the transcription of a spoken sentence by participants and the counting of the number of attempts required by the assistant to make the accurate transcription. In relation to the number of attempts, we assumed that a user would give up from being understood after a certain number of attempts. For the whole experiment, we considered that if after three attempts the transcription of the sentence was not accurate, the sentence would be classified as inaccurate. We selected for our experiments two smartphone assistants: Apple's Siri and Google Assistant. Twenty volunteers participated of the experiments: 10 of them interacted with Siri and 10 with Google. The profile of the participants varied in gender, nationality, accent, age, income range and education. Illiterate people were not included, but non-natives who could read portuguese and speak brazilian portuguese were.

Each participant was asked to read four sentences. The sentences contain words, chosen by HCI experts, that are often pronounced in the wrong way due to variation in pronunciation depending on the accent of the participants. For example, the word beneficent (in Portuguese "beneficente"), is commonly pronounced in the wrong way by people from certain regions of Brazil. Other words in which this occurs were used such as: "bicarbonato", "iogurte", "cabeleireiro", "problema", "padrasto", "cérebro", "entretido", "brócolis" and "crocante". In order to include a random factor in the process, seven sentences were elaborated and the participants made the draw of four of these seven. The sentences were elaborated as questions to more closely simulate the interaction with assistants. The seven sentences were the following:

"How do I get the privilege of going to a beneficent ("beneficente") event?"

"Is it bad to put baking soda ("bicarbonato") in yogurt ("Ãŕogurte")?"

"Where's the nearest hairdresser ("cabeleireiro")?"

"Where can I treat a thoracic problem ("problema")?"

"Is it coincidence meeting someone who has the same stepfather ("padrasto")?"

"What to do to make the brain ("cÃĭrebro") entertained ("entretido")?"

"How to cook the broccoli ("brócolis") to make it crispy ("crocante")?"

For each sentence read two metrics were collected, the accuracy of the sentence transcription and the number of attempts, the latter being limited to three attempts. All sessions were individual and conducted by two experts in Human Computer Interaction (HCI), during December 4-8th , 2018.

3.2 Characterization of the Participants

The twenty participants were divided into two groups of ten. In order to reduce learning effects it was conducted a Between-Subjects study [5], thereby each group interacted with only one of the assistants.

The participants profile containing their characteristics of gender and region were collected via a socioeconomic cultural questionnaire and anonymized.

The distribution of participants by gender and accent was not uniform. Out of the twenty participants seven were female and 13 male. Information about the accent was stated by the participant, usually this information was filled out with their place of birth, often the participants declared the city, state or country of origin, in the case of foreigners fluent in Brazilian Portuguese. As for the accent, native participants were aggregated by brazilian regions.

Eight participants declared an accent from the Northeast region, six from the Southeast region, two Foreigners who do not speak Portuguese natively and four of participants were aggregated as Others because of their lack of representativeness if they were considered by their proper regions, one was from the South, one from the Midwest and two were from the North .

Table 1: Distribution of Participants by Accent

Region	Participants	Proportions
Northeast	8	40%
Southeast	6	30%
Others	4	20%
Foreign	2	10%
Total	**20**	**100%**

3.3 Metrics of Evaluation

We defined two metrics to assess the quality of the recognition of the sentences spoken to the assistants. The first metric was the transcription, by the assistant, of the sentence spoken by the participant. Each transcript metric could assume two values: accurate and inaccurate. The label accurate was assigned in cases where the transcription of all words pronounced were done in correct Portuguese, even in cases where a word was pronounced incorrectly, for example, if the participant pronounced "beneficiente" and the assistant transcribed "beneficente" as should be in the formal Portuguese.

The label inaccurate was attributed to the transcripts wrongly done even in cases in which the assistant used a word similar to the one that was pronounced. For example, the assistant transcribed "yakult" (a brand of probiotic fermented milk) when the participant pronounced the word yogurt in the sentence.

The second metric is the number of attempts. It refers to the number of times the participant repeated the sentence if it was not transcribed exactly. There was a limitation of three attempts for each sentence, considering that in real-world interactions users have a tolerance of system errors until withdrawal. Thus, for example when the participant pronounced yogurt and the assistant transcribed "yakult" the participant could pronounce the sentence up to three times.

4 RESULTS

4.1 Number of Attempts and Quality of Transcriptions

Each of the twenty participants read four sentences, that should generate a base of 80 readings if all the readings had been accurate. As 24 sentences have been read a second time and 11 have been read a third time, the total base contains 115 readings. The proportion of accurate and inaccurate by tries of reading is presented in Figure 1. In the first round of readings (T1), 70% of the sentences have been transcribed accurately and 30%(24 sentences) in an inaccurate form. These 24 sentences were read again (T2) and 54%had an accurate transcription and 46%inaccurate. Finally, a third round of reading (T3) was performed with the eleven sentences that have been inaccurately transcripted. Since the maximum number of tries were limited to 3, six out them have not been transcripted correctly.

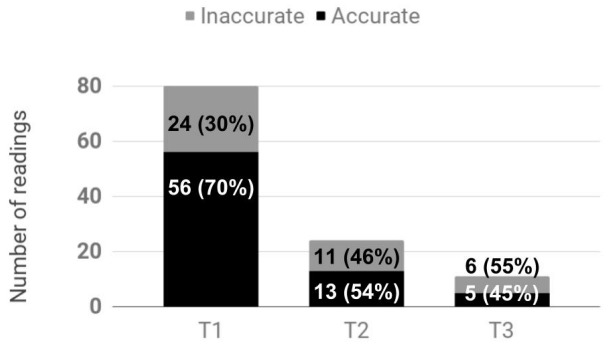


Figure 1: Proportion of accurate and inaccurate readings by attempts.

4.1.1 Siri vs Google. Figure 2 shows the proportion of accurate and inaccurate readings for Siri (left side) and Google Assistant (right side). The number of times that Google assistant transcribed , in the first attempt (T1), a sentence read by the participants were greater than Siri's one (88% vs. 52%). It is worth noting that even though Google assistant has a better performance in transcribing the sentences correctly in the first attempt, both reach similar levels of sentences without correct transcription (3 transcriptions of sentences were inexact for each assistant). A qualitative analysis showed that the sentence " Como faço para ter o privilégio de ir a um evento beneficente?" (sentence a in Section 3.1) has been wrongly transcribed by both assistants. The three sentences that Siri transcribed non-exactly were pronounced with an accent from Northeast while Google assistant had two sentences pronounced by non-native speakers and one with accent from the group Others.

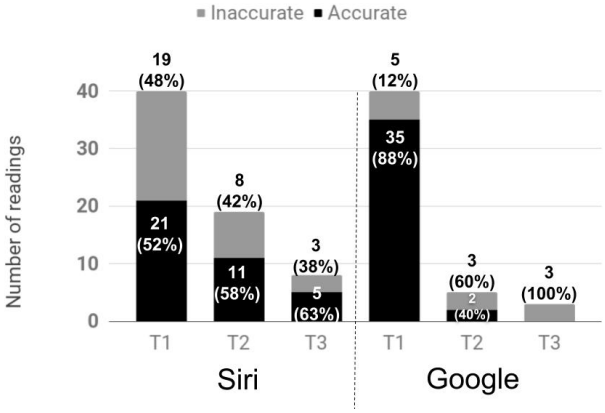


Figure 2: Proportion of accurate and inaccurate readings for Siri and Google.

4.1.2 Female vs Male. Female participants were better understood by assistants than male participants. Figure 3 shows the proportion of the distribution of accurate and inaccurate by tries of reading by female participants (left side) and male participants (right side).

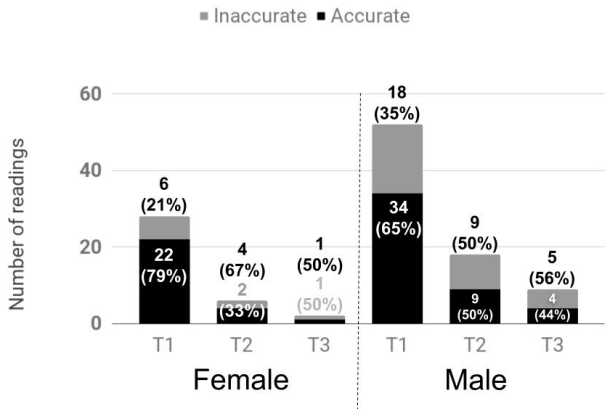


Figure 3: Proportion of accurate and inaccurate readings for Siri and Google.

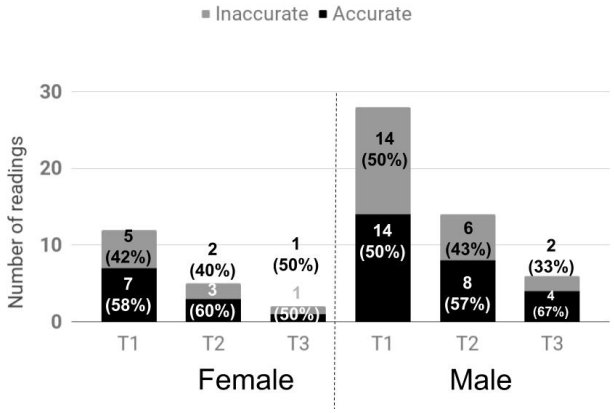


Figure 5: Proportion of accurate and inaccurate readings by female and male participants using Siri.

4.1.3 By Accent. Participants with an accent from the Southeast were better understood by the assistants than participants from other regions. Figure 4 shows the proportion of accurate and inaccurate for the tries of reading by participants, grouped by the region of accent: Northeast, Southeast, Others and Foreign.

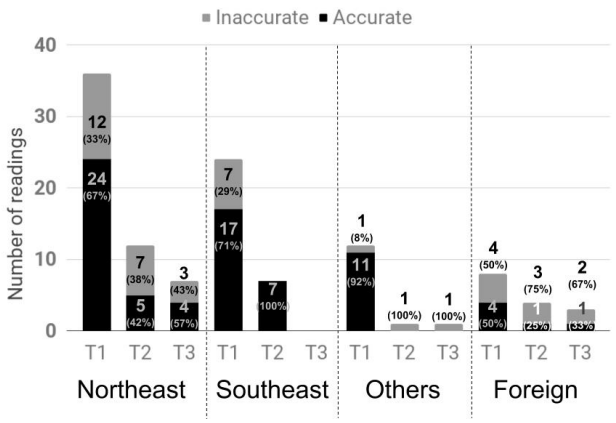


Figure 4: Proportion of accurate and inaccurate readings by region of the accent of the participants.

4.1.4 Siri by Gender. Siri had a better understanding of female participants than of male participants. Figure 5 shows the proportion of accurate and inaccurate for the tries of reading for Siri for female and male participants.

4.1.5 Google Assistant by Gender. Google also had a better understanding of female participants than of male participants. It was better than Siri for both female and male. Figure 6 shows the proportion of accurate and inaccurate by tries of reading for Google for female and for male participants.

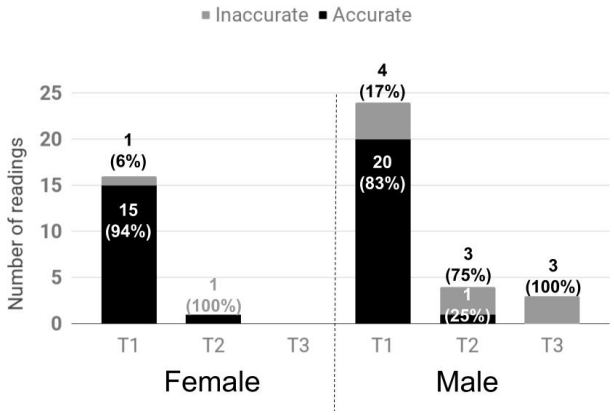


Figure 6: Proportion of accurate and inaccurate readings made by female and male participants using Google Assistant.

4.1.6 Siri by Accent. As aforementioned, Siri had a better understanding of participants of the Southeast region. Surprisingly it worked better even for Foreigners than for participants from the Northeast region. Figure 7 shows the proportion of accurate and inaccurate by tries of reading by participants from the different regions of accent. There were no participants from Others using Siri.

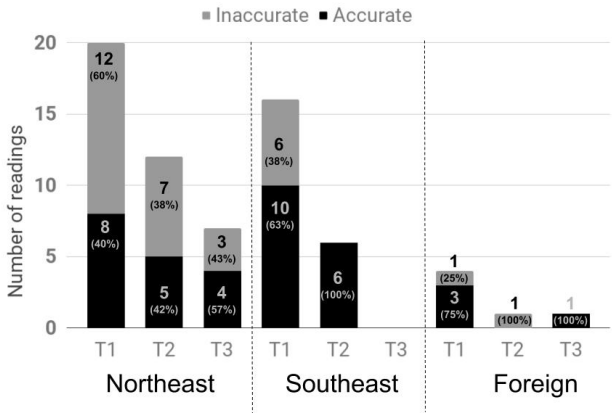


Figure 7: Proportion of accurate and inaccurate readings by region of the accent using Siri.

4.1.7 Google Assistant by Accent. Google had better understanding of participants of the Northeast. Figure 8 shows the proportion of accurate and inaccurate by tries of reading by participants from different regions.

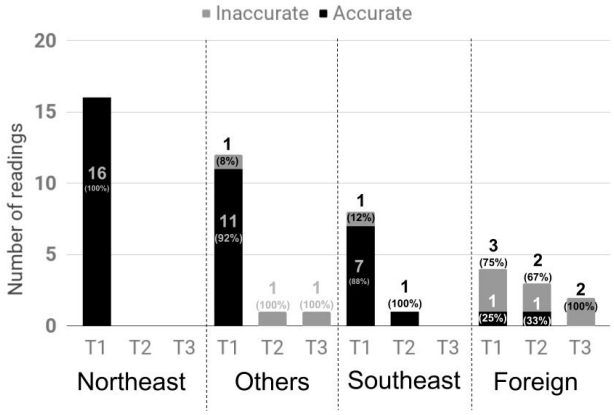


Figure 8: Proportion of accurate and inaccurate readings by accent using Google Assistant.

4.2 Summary

Table 2 presents the number of sentences read to each assistant, segmented by gender and accent. For instance, it shows that 12 sentences were read by female participants using Siri and 16 by female participants using Google. The arithmetic average of the number of attempts that the participants made to be understood (i.e., the number of tries to reach the exact transcription) is 1.50 for Siri and 1.06 for Google.

It evidences that, in average, the number of tries using Siri for participants with typical accent from the Northeast (2.00) is almost double the participants with accent from the Southeast (1.31). Considering gender segmentation, both assistants have better results for female participants, Google presented the best performance

Table 2: Number of sentences read and average of tries by assistant, Gender (Gn) and Accent (Ac). (S=Siri, G=Google), Accent: Northeast (NE), Southeast (SE), Others (O) and Foreign (F)

		Number of Sentences		Proportion of Accurate Transcriptions		Arithmetic Average of Tries	
		S	G	S	G	S	G
Gn	F	12	16	58%	94%	1.50	1.06
	M	28	24	54%	68%	1.71	1.21
Ac	NE	16	16	44%	100%	2.00	1.00
	SE	16	8	73%	89%	1.31	1.13
	O	4	12	-	79%	1.75	1.00
	F	4	4	67%	22%	1.50	2.25

when transcribing female readers, with an average of 1.06 tries.
We have applied the chi-square test to check whether transcriptions had a detectable difference regarding gender and accent depending on which assistant was used. The result of gender to Siri was χ^2(NA, N=66) = 1.02, p>0,05 and to Google χ^2(1, N=48) = 4.32, p<0.05 and of accent to Siri was χ^2(3, N=66) = 13.66, p<0.05 and to Google was χ^2(3, N=48) = 19.40, p<0.001. When comparing transcription between genders, there was a significant difference when the reading was made for Google Assistant. However, for Siri the difference in transcription regarding gender was not significant. Regarding the comparison between transcriptions by accent both Google Assistant and Siri presented a significant difference.

5 CONCLUSION

This study presents a preliminary analysis indicating that the training process of smartphone assistants, for the Brazilian Portuguese, can be biased towards voices of individuals from the most developed part of the country. It is particularly true for the case of Apple's assistant Siri when recognizing phrases with accent for the Southeast region. Variations in the quality of recognition on the basis of gender are also an indicator that further investigations need to be carried out in order to specify the reasons for it.
As future work, we plan to investigate if the results obtained here are also true regardless of whether the speech is composed of isolated words, short sentences, or long samples. We also plan to investigate variations on the transcriptions caused by the order of the sentences presented to the assistant, this can indicate whether the assistant has the potential to learn after unsuccessful tries. We consider that transcriptions are not the only way to evaluate the quality of the interaction via audio. The quality of a dialogue, for instance, depends on user's actions and choices. When an assistant fails to understand a word, an user tends to change (slightly or not) his/her intonation and/or timing, in the attempt to make the system understand the intended utterance. How much these changes bring prejudice to the dialog (e.g. making the dialog tedious) is a theme for further investigation.
Some limitations on the methodology of this preliminary study will be removed in the future, such as the unbalanced number of participants by gender and accent. We should construct the size

of the samples proportionally to demographic indicators of the country. Semantics was not a variable analyzed in the scope of this study, which would encompass regionalisms and words employed by specific social classes.

This work provides useful information for the definition of rules and policies regarding the evolution of voice-based assistants and devices in the developing world. Such policies must protect vulnerable social groups, that are excluded due to social, economic or regional issues.

REFERENCES

[1] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[2] Jerome R Bellegarda. 2014. Spoken language understanding for natural interaction: The siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 3–14.

[3] Estado de São Paulo Newspaper. 2017. WhatsApp chega a 120 milhões de usuarios no Brasil.

[4] Adriana de Souza e Silva, Daniel M Sutko, Fernando A Salis, and Claudio de Souza e Silva. 2011. Mobile phone appropriation in the favelas of Rio de Janeiro, Brazil. *New Media & Society* 13, 3 (2011), 411–426.

[5] Anthony G Greenwald. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83, 2 (1976), 314.

[6] Reuters Institute. 2018. The Future of Voice and the Implications for News.

[7] The Intercept. 2018. Amazon's accent recognition technology could tell the government where youâĂŹre from.

[8] Wired Magazine. 2017. Voice Is the Next Big Platform, Unless You Have an Accent.

[9] Rami Matarneh, Svitlana Maksymova, Vyacheslav Lyashenko, and N Belova. 2017. Speech recognition systems: A comparative review. (2017).

[10] Tyler K Perrachione. 2017. Speaker recognition across languages. Oxford University Press.

[11] Moirangthem Tiken Singh, Abdur Razzaq Fayjie, and Biswajeet Kachari. 2015. Speech Recognition System for North-East Indian Accent. In *International Journal of Applied Information Systems (IJAIS)*. Vol. 9. Citeseer.

[12] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Daniel Jurafsky, Rebecca Starr, and Su-Youn Yoon. 2005. Accent detection and speech recognition for shanghai-accented mandarin. In *Ninth European Conference on Speech Communication and Technology*.

Managing Bias in AI

Drew Roselli
ParallelM
Sunnyvale, CA, USA,
drew.roselli@parallelm.com

Jeanna Matthews
Department of Computer Science
Clarkson University, Potsdam, NY,
USA, jnm@clarkson.edu

Nisha Talagala
Saratoga
CA, nishatalagala@gmail.com

ABSTRACT

Recent awareness of the impacts of bias in AI algorithms raises the risk for companies to deploy such algorithms, especially because the algorithms may not be explainable in the same way that non-AI algorithms are. Even with careful review of the algorithms and data sets, it may not be possible to delete all unwanted bias, particularly because AI systems learn from historical data, which encodes historical biases. In this paper, we propose a set of processes that companies can use to mitigate and manage three general classes of bias: those related to mapping the business intent into the AI implementation, those that arise due to the distribution of samples used for training, and those that are present in individual input samples. While there may be no simple or complete solution to this issue, best practices can be used to reduce the effects of bias on algorithmic outcomes.

CCS CONCEPTS

• **Computing methodologies~Artificial intelligence** • **Computing methodologies~Machine learning** • Security and privacy~Social aspects of security and privacy

KEYWORDS

Artificial intelligence; bias; production monitoring

ACM Reference format:

Drew Roselli, Jeanna Matthews and Nisha Talagala. 2019. Managing Bias in AI. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317590

1. Introduction

The potential for efficiencies and improvements has attracted many companies to invest in AI [1] [2]. However, the specter of unintentional bias and its potential to cause harm can negatively impact a company's reputation. This type of risk and consequent exposure to lawsuits is an important reason for caution. Bias in algorithms not only carries these risks, it can cause an application to perform sub-optimally, leading to missed opportunities. For example, bias in lending practices can lead to both financial loss, when bias favors some people, and missed financial gain when bias unfairly discriminates against others. The situation is exacerbated by the fact that many AI systems are not "explainable" in ways that deployers can claim to not contain unintended bias [3] [4] and other categories of errors. Managers have a difficult time trusting key business decisions to inferences that cannot be explained, especially when bias and other sub-optimal decision making patterns can be embedded within data used for training in non-obvious ways [5] [6]. Even worse, if the output of the application can impact its input, it can maintain a vicious cycle of bias that mires the application in a realm of poor predictive performance indefinitely [7].

With so much at risk, it is time to develop a set of best practices to help guide deployers of AI systems around these potential pitfalls. As a first step in this direction, we propose a framework to understand, quantify, and mitigate sources of bias. First, we survey the potential sources and impacts of bias in AI systems. Then we describe processes that can help companies manage bias to improve their AI outcomes and help provide confidence in the results for themselves, their customers, and regulators.

2. Potential Problems

Bias in AI systems comes from diverse sources, from the algorithm and input attributes chosen to the hidden correlations within training data. We describe three general classes of bias: those related to mapping the business intent into the AI implementation, those that arise due to the distribution of samples used for training (including historical effects), and those that are present in individual input samples.

2.1. Issues with Representing the Goal

The earliest stage for bias to creep into an AI system is when the deployer determines its concrete algorithmic objective from what may be a nebulous goal. For example, the actual goal of a business may be to direct advertising to the potential customers most likely to purchase their product. Since there is no straightforward way to map this into an AI implementation [8], companies must choose the hypothesis, input attributes, and training labels or reinforcement criteria that they deem will best accomplish this goal. For example, a company selling video games might hypothesize that their product would be most likely to sell to young men and look for customers with attributes such as male and aged 15-25.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317590

2.1.1. Proxy Goals. In the example above, since there is no way to determine the exact likelihood of someone to purchase their product, they may choose a goal of selecting individuals with attributes similar to those of customers that previously purchased a similar product. However, this may not be the right choice when trying to enter new markets or trying to market new product features. Basing the goal choice on historical information without factoring in suitable context will necessarily expose the system to historical bias. For example, people from some regions may simply not have purchased the product because it wasn't previously marketed in their region. Similarly, changes in product features, price, or external trends may render it suitable for customers who had not purchased the previous version.

2.1.2. Feature Selection. The choice of which attributes to include is perhaps the most obvious source of bias from the mapping designers. For example, a university admission system could look at standardized test scores, rank in class, GPA, letters of recommendation, etc. While the end goal may be the same (i.e., predicting success in college), the choice of features used can result in very different decisions. Even seemingly innocuous input features can contain hidden biases [5, 9]. Even more difficult to quantify is bias that arises from features that are not included but could favorably influence the predictions for some people if included. Additional information, such as personal observations from a letter of recommendation, may be harder to map to quantifiable and well-defined features, rendering such useful information unavailable to the algorithm.

2.1.3. Surrogate Data. AI models demand large sets of mathematical input. Training sets must be represented numerically, and because the training sets must be large, these features are limited to those that developers can easily acquire at scale. This may mean that a job screening service uses credit scores rather than letters of recommendation as a surrogate for a feature such as "reliability" [29]. Such mathematical reductions can cause information loss which then biases the problem mapping. Surrogate data can also introduce bias by serving as proxies for restricted input, for example, zip codes can be used as proxy for race, magazine subscriptions for race or gender, and purchasing patterns for medical conditions.

2.2. Issues with Data Sets

In addition to dataset issues that can create problems during the mapping phase, there may be issues with training or production datasets. Creating training sets can be an arduous task. It typically involves preening a large data set [10] and for supervised learning, it includes obtaining labels. For deep learning systems, it may include ensuring rare cases are proportionally overrepresented to give the model adequate opportunity to learn such cases during training. Due to the scale, complexity, and sometimes timeliness of this task, creating a training data set can be the bulk of the effort required in AI systems and is often the source of problems [11]. Training datasets can also be manipulated, rendering the AI algorithm vulnerable [13].

2.2.1. Unseen Cases. Much of the advantage of AI systems are their ability to generalize solutions with robustness to varied input. However, this can become a disadvantage when the system is faced with a class for which it was not trained. For example, a neural network trained to classify texts as German or English will still provide an answer when given a text in French rather than saying "I don't know". Such issues can lead to "hidden" or "silent" mispredictions, which can then propagate to cause additional harm to the business application.

2.2.2. Mismatched Data Sets. If data seen in production differs significantly from that used in training, the model is unlikely to perform well. Extending the point above, commercial facial recognition systems trained on mostly fair-skinned subjects have vastly different accuracies for different populations: 0.8% for lighter-skinned men and 34.7% for darker-skinned women [30]. Even if the model is originally trained on a dataset that matches production use, production data can change over time due to various effects from seasonal changes to external trigger events. Any such change can bring about hidden effects generated by mismatched data sets.

2.2.3. Manipulated Data. Training data can be manipulated to skew the results as was exemplified by the short-lived chatbot Tay, which quickly mimicked the hate speech of its Twitter correspondents [12, 31]. Systems trained on small, public data sets are especially vulnerable to this form of attack. Similarly, data poisoning is a known security challenge for AI systems [13].

2.2.4. Unlearned Cases. Even well-trained models do not have 100% accuracy; indeed, such high accuracy would likely result from overfitting the data and indicate that the model is not likely to generalize well to new cases. As a result, even well-trained models will have classes of samples for which they perform poorly. Studies have shown that facial recognition datasets that do not adequately represent across ethnic groups can cause trained models to display vastly different accuracies across race [14].

2.2.5. Non-Generalizable Features. Due to the practical difficulties in creating large, labeled training sets, model developers may rely on training from well-preened subsets of their expected production data sets. This can result in granting importance to features that are particular to the training set and not generalizable to broader data sets. For example, [15] shows how text classifiers, which were trained to classify articles as "Christian" or "atheist" on standard newsgroup training sets, emphasize non-relevant words like "POST" in making their classifications due to the distributions of those words in the training set.

2.2.6. Irrelevant Correlations. If the training data contains correlations between irrelevant input features and the result, it may produce incorrect predictions as a result. For example, Ribeiro et al. trained a classifier to differentiate between wolves and dogs with images of wolves surrounded by snow and dogs without snow. After training, the model sometimes predicts that a dog surrounded by snow is a wolf [15]. Unlike non-generalizable features, the distribution of irrelevant correlations may not be particular to the training set but may occur in real-world data as well. It may well be that wolves are more likely to be found in snow than dogs. However, it would be incorrect for the feature to impact the prediction; a wolf is still a wolf even when it is in Grandmother's house.

2.2.7. Issues with Using Historical Data. AI systems necessarily learn from past information. Unfortunately, this includes learning human biases contained therein [5] and potentially missing opportunities that emerge from changing environments [9].

2.3. Issues with Individual Samples

We classify issues with individual samples as those that can be seen by examining the data of a single sample. The problem may be with just that sample or endemic to all the samples. This classification is important when the data sets contain personal information such that the entire data set cannot be made publicly viewable, but where individuals may be able to review their own personal data.

2.3.1. Inaccurate Data. Data used for training is often highly curated in order to ensure the model learns effectively. Unfortunately, real-world data is seldom so clean. It can be incomplete or corrupted. For data that is manually entered, it can be entered incorrectly [27]. Data that is automatically collected can have incorrect sources [28].

2.3.2. Stale Data. Data used for both training and production input can be out-of-date. This may particularly be the case when large "dictionaries" are cached for fast access time. For example, credit scores could be downloaded from an external source and stored locally for fast access. Unfortunately, there may be resistance to updating the dataset by developers as it may reset the baseline for ongoing training experiments.

3. Managing Bias

Given the number of potential sources of bias currently known (and more are being discovered as the field matures), it can be daunting to face how to tackle them. Considering the myriad of issues that cause bias to enter the system, we cannot expect a single approach to resolve all of them. Instead, we propose a combination of quantitative assessments, business processes, monitoring, data review, evaluations, and controlled experiments. Before detailing the above stages, we establish some ground rules for processes we want to include.

First, any process to evaluate an AI system for bias must be achievable by those that are not necessarily the primary developer. This is important because the system may need to be understood by non-technical management or evaluated by an auditor or regulator. In fact, since complex models are increasingly available from external sources [16, 17], even the primary developer may not know the model's inner workings. Therefore, we require that our processes not require any knowledge of the internal workings of the model itself; we treat the entire AI pipeline (including any feature engineering) as a black box and use only the input and output data for evaluation.

Second, transparency of input data is important. This is the only way to verify that the data is accurate and does not contain protected or incorrect data. Even if the data is private, it should be viewable by the person about whom the data is [18]. Methodologies for data versioning, data cataloging, and data tracking and governance are also critical to ensure that the specific dataset used to train a given model can always be identified and examined.

We group the processes into stages at which the process most prominently come into play during the deployment lifecycle of the AI system. However, this is mostly an organizing technique for planning these processes as the stages are likely to overlap and iterate over the lifetime of the application.

3.1. Substantiate Assumptions

During the planning stage of the application, prepare to provide quantitative evidence for the validity of your chosen numerical representations, the hypothesis itself, and the impact of the application on its environment, including its future input.

3.1.1. Substantiate Surrogate Data. When surrogate data is used, it should be accompanied by quantitative evidence that suggests that the surrogate data is appropriate for its intended use in the model. Known limitations of the surrogate data should be documented and presented during reviews of predictions [19].

3.1.2. Substantiate the Hypothesis. Similarly, the model's intended use and appropriateness for that use should be accompanied by quantitative external evidence that supports it along with any known limitations to the methodology [19].

3.2. Vet Training Data

Training data needs to be vetted for accuracy, relevance, and freshness. Since training data is often curated, it tends to be "cleaner" than inference data. For example, incomplete or ambiguous samples may be eliminated from the set. Unfortunately, the effort required to curate the training data may be at odds with keeping it fresh. On the other hand, production inference data is likely to be fresh but may contain samples that are incomplete or ambiguous.

3.2.1. Avoid Overly Curated Training Data. When considering whether the training set is appropriate for the expected production workload, avoid selecting highly curated sets that perform well on curated evaluation data but are unlikely to perform well in production.

3.2.2. Guard Against Manipulation. Consider whether the training data can be manipulated by outside (or inside) actors. If manipulation cannot be prevented, consider alternative data sources. Suciu et al. present a helpful overview of literature on poisoning machine learning in the context of a generalized model for the capabilities of adversarial agents [20]. Data security practices may also be needed to ensure that inappropriate datasets are not accidentally accessed by those building AI models. In large organizations where teams of data scientists work, such access restrictions can be critical.

3.3. Evaluate for Bias

As part of your model's evaluation, include the following in addition to standard accuracy metrics.

3.3.1. Utilize Targeted Toolsets. With the growing recognition of issues of bias in AI, there are an increasing number of tools available to help detect and mitigate bias in attribute sets, feature engineering, or the model itself [21].

3.3.2. Validate Predictions. Although we treat AI systems as a black box, there are tools available that can determine which

input features determined the resulting prediction [15, 22]. Such predictions can be validated by providing a qualified judge (who does not a priori know the AI system's prediction) with the same feature values and comparing the judge's determination with the AI's prediction. This both validates and provides explanations for predictions. Such an evaluation process can help identify both non-generalizable features and irrelevant correlations.

3.3.3. Match Training to Production Needs. It is important to monitor the distribution of input data to see whether production data is consistent with the data used in training. Input data should be monitored for anomalous cases that differ substantially from those seen in training. Input streams should also be monitored to ensure that the distribution of data seen in production does not stray from the distribution anticipated by the training data set.

3.4. Monitor Production Data

AI accuracy is based on its evaluation data set. When the input seen in production does not match this, the system's performance should be flagged and treated with skepticism.

3.4.1. Detect Incomplete Data. Production AI systems should actively monitor for incomplete data. This is particularly true when the system is trained on cleaned data sets but then receives a wider range of samples in production. For example, an employer recruiting program may use credit scores to train a filter to screen candidates and confuse someone with no credit history as someone with a low credit history.

3.4.2. Detect Data Divergence. Production input monitoring should evaluate whether the input data received for inferences matches the expected distribution anticipated by the training data. If not, the model should be re-trained with fresher data to match the production workload [23].

Note that this is similar to the need to match training data to production needs that we discussed earlier. The first occurs during the planning of the training set and the second occurs in monitoring production data to detect when assumptions made during the training phase no longer hold.

3.4.3. Detect Uncommon Cases. Inputs with unique feature distributions or low confidence should be detected and ideally flagged for external labeling and inclusion into future training sets. Chen et al. have shown that supplementing training sets with additional data can be an effective strategy for reducing bias [24].

3.4.4. Refresh Input Data Regularly. Input data sources should be refreshed in a timely manner. This includes data used for training as well as any stored data sets that are included as features for production inference.

3.5. Create Supportive Processes

In addition to the data science and engineering aspects of managing bias, there also needs to be business practices in place to enable reviews and emphasize testing new directions.

3.5.1. Enable Input Data Review and Correction. Individuals should be able to view data that is about them personally. There should be a process in place for them to challenge and correct this data [18]. Ideally, there should also be processes for individuals to provide alternate information that directly speaks to the goal (and not just the proxy goals) of the system and have that additional information considered by a human reviewer.

3.5.2. Create Mechanisms to Overcome Historical Bias. Adding some randomization to outcomes can be used to avoid consistent discrimination from historical data [25]. Similarly, A/B experiments can be used that drop some features. Ideally, the results of such predictions can be fed back into the system to produce continuous improvement.

3.6. Quantify Feedback Loops

In cases for which the AI predictions can impact its own future input, it is important to ensure that this impact is quantified by comparing against a suitable control or comparing the results against external evidence. For example, algorithms that predict crime based on police reports may cause more police to be deployed to the site of the predictions, which in turn may result in more police reports [7]. Note that not all feedback loops are considered negative; algorithms that return search results may return better search results when less common search terms are used, which may result in users learning to enter less common search terms. However, even in positive cases, it is important for the designers to understand the impact of feedback loops.

4. Related Work

While there exists robust literature of potential problems with bias in AI systems, there is considerably less available for how to manage this bias, especially across the wide spectrum of potential causes of bias. A number of principles documents and primers for algorithmic accountability have been published by researchers [32] and groups such as ACM [33, 34], World Wide Web Foundation [35] AI Now [36], and Data and Society [37]. Other researchers have proposed taxonomies of bias that serve as an excellent starting point towards managing that bias [38, 39, 40] Cramer et. al describe their experience with translating these principles documents and taxonomies into concrete processes within their organization, Spotify [26]. We similarly share the goal of mapping research literature into practical processes for industry teams and focus on processes that can be integrated into existing product cycles including substantiating surrogate data, vetting training data, and validating predictions. Cramer et al. point out that in early product development data sets may be limited and thus teams may be making decisions in the presence of skewed information. Some of the techniques we propose for detecting data divergence can help with this by actively monitoring and reacting when production data varies substantially from the assumptions made during training.

5. Summary

Many data sets contain bias that inevitably impacts AI systems that use that data. Even careful cleansing of protected and proxy data will not completely remove the results of historical bias contained within the data. However, putting in place a combination of processes can mitigate the impact of bias on current pre-

dictions and reduce its impact over time. We advocate a combination of processes:

Table 1: Issues that cause bias and recommended mitigation actions.

Issue type	Issue	Mitigations
Representing the problem	Proxy Goals	• Substantiate the hypothesis with external quantitative evidence. • Evaluate impact of predictions against external metrics.
	Feature Selection	• Utilize targeted toolsets to help identify hidden bias. • Validate prediction reasons as part of model evaluation. • Evaluate unlearned cases for clues to missing attributes.
	Surrogate Data	• Substantiate surrogate data noting known limitations. • Enable an input data review process to help catch non-obvious limitations.
Datasets	Unseen Cases	• Detect and review unexpected input feature patterns.
	Mismatched Data Sets	• Avoid overly curated training data. • Detect when input data diverges from distributions anticipated during training.
	Manipulated Data	• Consider how training data and data used as input could be manipulated and enable safeguards or use alternative sources.
	Unlearned Cases	• Review the cases that were not learned during training as part of model evaluation.
	Non-generalizable Features	• During model evaluation, determine which features were used to make predictions and compare them against those used by an external judge.
	Irrelevant Correlations	
	Issues with Historical Data	• Utilized targeted toolsets that help remove historical bias. • Supplement training sets with under-represented samples. • Utilize randomization and A/B experiments to explore new outcomes.
Individual Samples	Inaccurate Data	• Detect incomplete data. • Enable an input data review process to help catch and correct incorrect input.
	Stale Data	• Refresh cached data sets regularly.

- Data monitoring to detect anomalies and determine whether the training set is appropriate.
- Quantitative analysis to justify surrogate data and account for feedback loops.

- A review process that ensures input correctness.
- An evaluation process that verifies the predictions are grounded in reasonable feature values.
- Mechanisms to explore new outcomes and incorporate these outcomes into future predictions to overcome historical bias.

Table 1 maps the problems described in Section 2 to the mitigation actions described in Section 3.

Reviews of AI systems for bias should ensure all of the above are implemented. As AI systems become more competitive, mitigating bias may become more necessary not only to avoid liability but to explore opportunities that may otherwise be missed. As awareness of the potential harmful impacts of bias in AI algorithms grows, we can encourage companies to mitigate the bias in AI systems by emphasizing the benefits of doing so such as increasing competitiveness and avoiding hazards and by outlining practical management processes that can be followed to actively identify and reduce that bias.

REFERENCES

[1] Witten, H., Frank, E., and Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA.

[2] McKinsey Global Institute (2017). "Artificial Intelligence: The Next Digital Frontier?"

[3] Goodman, B. and Flaxman, S. (2016). "European Union regulations on algorithmic decision-making and a 'right to explanation'". ICML Workshop on Human Interpretability in Machine Learning, New York, NY.

[4] Kahn, J. (2018). "Artificial Intelligence Has Some Explaining to Do". *Bloomberg Businessweek*, 12 Dec 2018.

[5] Bolukbasi. T., Chang, K., Zou, J., Saligrama, A., and Kalai, A. (2016). "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". *Proceedings of the 30th International Conference on Neural Information Processing Systems*.Barcelona, Spain.

[6] Angwin, J., Larson, J, Mattu, S., and Kirchner, L. (2016). "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks". ProPublica, 23 May 2016.

[7] Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018). "Runaway Feedback Loops in Predictive Policing". *Proceedings of Machine Learning Research*.

[8] Hao, K. (2019). "This is How AI Bias Really Happens -— and Why It's so Hard to Fix". MIT Technology Review, 4 Feb 2019.

[9] Dastin, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women". Reuters Business News, 10 Oct 2018.

[10] Del Balso, M. and Hermann, J. (2017). "Meet Michelangelo: Uber's Machine Learning Platform". Uber Engineering, 5 Sep 2017.

[11] Baylor. D. et al. (2017). "TFX: A TensorFlow-Based Production-Scale Machine Learning Platform". *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Nova Scotia, Canada.

[12] Vincent, J. (2016). "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day". The Verge, 24 Mar 2016.

[13] Bursztein, E. (2018). "Attacks against machine learning - an overview". https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/.

[14] Buolamwini, J. (2019). "Aritificial Intelligence Has a Problem with Gender and Racial Bias". TIME, 7 Feb 2019.

[15] Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "'Why Should I Trust You?': Explaining the Predictions of Any Classifier". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA.

[16] Google. "TensorFlow Hub," https://www.tensorflow.org/hub.

[17] H2O. "Open Source Leader in AI and ML". h2o, https://www.h2o.ai/.

[18] European Commission. "Rights for Citizens - European Commission". https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens_en.

[19] Board of Governors of the Federal Reserve System (2011). "The Fed - Supervisory Letter SR 11-7 on Guidance on Model Risk Management". https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

[20] Suciu, O., Marginean, R., Kaya, Y., Daume. H. III, and Dumitras, T. (2018). "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks".*27th USENIX Security Symposium*, Baltimore, MD.

[21] IBM. "AI Fairness 360". http://aif360.mybluemix.net/.

[22] H2O. "Interpreting a Model — Using Driverless AI". https://www.h2o.ai/wp-content/uploads/2017/09/driverlessai/interpreting.html.

[23] Ghanta, S. and Talagala N. (2018). "MLOps Health: Taking the Pulse of ML in Production". ITOpsTimes, 10 July 2018.

[24] Chen, I.Y., Johansson, F.D., and Sontag, D. (2018). "Why Is My Classifier Discriminatory?".*32nd Conference on Neural Information Processing Systems*, Montreal, Canada.

[25] Kroll, J., Huey, J., Barocas, S, Felten, E., Reidenberg, J., Robinson, D., and Yu, H. (2017). "Accountable Algorithms". *University of Pennsylvania Law Review*, vol. 165, pp. 633-705.

[26] Cramer, H., Garcia-Gathright, J., Springer, A, and Reddy, S. (2018). "Assessing and addressing algorithmic bias in practice". *ACM Interactions*, vol. XXV, no. 6, pp. 58-63.

[27] Bier, D. (2017). "E-Verify Has Delayed or Cost Half a Million Jobs for Legal Workers". CATO Institute, 16 May 2017. https://www.cato.org/blog/e-verify-has-held-or-cost-jobs-half-million-legal-workers.

[28] Suhartono, H., Levin, A., and Johnsson, J. (2018). "Why Did Lion Air Flight 610 Crash? New Report Details Struggle". Bloomberg, 27 Nov 2018. https://www.bloomberg.com/news/articles/2018-11-27/lion-air-pilots-struggle-detailed-in-preliminary-crash-report.

[29] 0'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown. New York, NY.

[30] Buolamwini, J. and Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, New York, NY.

[31] Buranyi, S. (2017). "Rise of the racist robots - how AI is learning all our worse impulses". The Guardian, 8 Aug 2017. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

[32] Nicholas Diakopoulos and Sorelle Friedler (2016). How to Hold Algorithms Accountable, MIT Technology Review, November 17, 2016. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/

[33] Garfinkel, S., Matthews, J., Shapiro, S., and Smith, J. (2017). Toward Algorithmic Transparency and Accountability. Communications of the ACM. Vol. 60, No. 9, Page 5, Sept. 2017, 10.1145/3125780.

[34] ACM US Public Policy Council (USACM) (2017). Statement on Algorithmic Transparency and Accountability. January 12, 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[35] World Wide Web Foundation (2017). Algorithmic accountability report, 2017. https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf

[36] Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now institute. Apr. 2018. https://ainowinstitute.org/aiareport2018.pdf

[37] Caplan, R., Donovan, J., Hanson, L., and Matthews, J. (2018). Algorithmic accountability primer. Data & Society. Apr. 2018. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

[38] Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. ACM Trans. Inf. Syst. 14, 3, 330–347.

[39] Olteanu, A., Castillo, C, Diaz, F., and Kiciman, E. (2016). Social data: Biases, methodological pitfalls, and ethical boundaries.http://dx.doi.org/10.2139/ssrn.2886526

[40] Baeza-Yates, R. (2016). Data and algorithmic bias in the web. Proceedings of the 8th ACM Conference on Web Science. ACM. New York. https://doi.org/10.1145/2908131.2908135

Uncovering Social Media Bots: a Transparency-focused Approach

Eric F. Santos
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro
eric.ferreira@ppgi.ufrj.br

Danilo S. Carvalho
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro
dsc@ufrj.br

Livia Ruback
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro
liviaruback@gmail.com

Jonice Oliveira
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro
jonice@dcc.ufrj.br

ABSTRACT

As the Online Social Networks (OSNs) presence continues to grow as a form of mass communication, tensions regarding their usage and perception by different social groups are reaching a turning point. The number of messages that are exchanged between users in these environments are vast and brought a trust problem, where it is difficult to know if the information is from a real person and if what was said is true. Automated users (bots) are part of this issue, as they may be used to spread false and/or harmful messages through an OSN while pretending to be a person. New attempts to automatically identify bots are in constant development, but so are the mechanisms to elude detection. We believe that teaching the user to identify a bot message is an important step in maintaining the credibility of content on social media. In this study, we developed an analysis tool, based on media literacy considerations, that helps the ordinary user to recognize a bot message using only textual features. Instead of simply classifying a user as a bot or human, this tool presents an interpretable reasoning path that helps to educate the user into recognizing suspicious activity. Experimental evaluation is conducted to test the tool's primary effectiveness (classification) and results are presented. The secondary effectiveness (interpretability) is discussed in qualitative terms.

KEYWORDS

online social network; bot message detection; media literacy

ACM Reference Format:
Eric F. Santos, Danilo S. Carvalho, Livia Ruback, and Jonice Oliveira. 2019. Uncovering Social Media Bots: a Transparency-focused Approach. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3317599

1 INTRODUCTION

Online Social Networks (OSNs) became a world changing trend on communications in the recent years. This kind of media allows different people, from different countries to share their lives and thoughts. The interactions between people on this environment brought a problem of trust, where it is difficult to know if another user is a real person and if what was said is true. Automated users (*bots*) are part of the issue, as they may be used to spread false and/or harmful messages through an OSN at a fast rate, contaminating real users with misinformation.

Not all the automated users are malignant, for instance, there are bots that spread messages from newspapers and weather applications as an automatic service to publish the same news in various channels. However, OSNs often have been dealing with malignant users and their activities. The most common malicious activity is spamming, wherein an automated user (bot user) disseminates content or malware/viruses to users of the social networks [4]. These bots can be used for several proposes, including: (1) advertising; (2) promoting politically oriented views and opinions; (3) promoting financial trends; (4) generating product reviews; (5) spreading malware, spam, and harmful links; (6) influencing search engine results such that particular links are shown first; (7) generating news feeds; and (8) creating an underground marketplace for purchasing social media followers [1].

The increasingly use of bots proliferating their biased messages can have a negative impact on Society, interfering with the people's democratic, civil and behavioral process. As an example we can mention what occurred in the United States of America presidential election of 2016. Studies suggests that so called "fake news" (i.e., false or misleading statements) might have been decisive to the victory of the current president on the election [5], and that such misinformation was spread by bots [16] on OSNs. Twitter, a widely used OSN where users can share short messages with text, images and videos, admitted to having excluded more than 50,000 bot accounts related to fake news propagation in US election [18].

Another problem that contribute to the misinformation dissemination is that regular (human) users often spread messages without checking whether they are true or not. Such users become new channels for misinformation, filtering and directing it according to their own ideological leanings.

Taking these problems into account, we believe that teaching the user to identify a bot message is an important step towards a healthier environment in social media. We conducted a study to analyze textual features taken from bot and human messages, used in previous related studies. We verify the feasibility for fast bot

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317599

detection while producing a useful guidance for users to identify a bot message. Such guidance is important from a media literacy standpoint, so that users can keep trust and gain more knowledge from their interactions and OSNs can guarantee credibility in the transmitted contents. We also provide a tool that performs a human/bot classification task and presents an interpretable reasoning path that helps educating users into recognizing suspicious activity.

Works approaching bots detection on OSNs typically collect many features, such as message content, network, profile, and others to characterize an automated user. However, combining such features may be resource intensive and also OSN specific. We focus on the text because it is the simplest way for a user to be aware that there may be something wrong with the message. The combination of many characteristics, such as number of followers, number of friends, among others, would make more difficult the literacy process, i.e., for a human to check all possibilities and judge if it was generated automatically.

Our solution is based on Twitter messages. Twitter is microblogging, a service that allows the user to share messages containing up to 140 characters (recently changed to 280), images, and videos. The user can play two roles: follower and followed. As a follower, the user can select another person which have an affinity and track their activities. As a followed, it shares your activities to other users that desire to follow your profile [11]. Recent works indicate a massive use of Twitter around the world. Aslan [3] has shown that there are about 500 million messages sent per day and 100 million daily active users. The message exchanges on Twitter achieved the mark of 500 million in the last year (about 5787 messages every second) [6], and 326 million people use this OSN every day. These numbers point to the platform's profile being focused for fast message creation and dissemination.

Experiments were conducted to evaluate the fitness of the proposed approach to the purposes of *(a)* identifying bots and *(b)* educating the user. The former was tested for classification performance (quantitative) and the latter for interpretability (qualitative).

The remainder of this paper is organized as follows: Section 2 introduces basic concepts and presents related works. Section 3 presents the approach to classify bots and humans and to assist users to distinguish between them. Section 4 details the experiments and discusses the evaluation results. Section 5 presents the guide tool, and Section 6 brings conclusions and final remarks.

2 BASIC CONCEPTS AND RELATED STUDIES

In this section, we summarize the main subjects covered in this work and the related studies in their respective areas: bot detection and media literacy. Bot detection research here focuses mainly into message features, and media literacy studies introduce the importance of this area.

2.1 Bot Detection

Detection of bots in OSNs as a task consists in classifying any given OSN user into bot or non-bot (human). As new approaches for detection are developed, so are bot countermeasures to avoid detection. Therefore, this task has no permanent solution.

Many of the works on bot detection collect many features, such as content, network, profile, and others to characterize an automated user. However, Martinez-Romo and Araujo [13] state that most of these features, such as the number of followers and friends, account creation dates, and others, can not only be easily manipulated by bots but that collecting all this data is a resource-intensive task. The research works presented in this section focus on the bot identification through the message, which is the common resource from most OSNs.

Martinez-Romo and Araujo [13] proposed an approach not to identify bot users, but spam messages. It applied statistical analysis of language to detect a spam message in Twitter trending topics. A statistical language model (SLM) is a probability distribution $P(s)$ over strings s that attempts to reflect how frequently a string s occurs in a sentence. The authors introduced an architecture that collects trending topics from the Twitter API [1], labels the messages, extracts the features, trains a classifier and detects a spam message. The collected trending topics date from 30 April 2012 to 8 May 2012, from English speakers. In order to label spam, messages were selected if the text contained a link. These links were classified using services that provide blacklisted sites. If a link were found in any blacklist, then the message was labeled as spam.

Igawa et. al. [10] studied a wavelet-based approach for account classification using only text messages generated by users in an online social network. This approach worked in conjunction with a new weighting scheme, called Lexicon Based Coefficient Attenuation (LBCA), that serves as input to a classifier algorithm. This research evidenced the low computational costs in identifying which kind of account will be analyzed since only the text is being taken into account. They use information retrieval techniques to analyze text content and conducted to conduct two experiments with Twitter datasets: matching the accounts with humans and bots, and identifying an account as human or bot. They used wavelets to decompose the signals brought by the weighting process and these signals are part of the detection process.

The dataset was collected from Twitter and only messages related to FIFA World Cup 2014 were retrieved, particularly the query "BRASIL" (Brazil, in Portuguese), "COPA" (World Cup, in Portuguese), "COPA2014", manually labeled. It used random forests and multilayer perceptron as classification algorithms and used other weighting schemes. However, the proposed LBCA had the best precision in both experiments using random forest classifier. In a later exploration, it proposed to analyze the behavior of the content produced by bots for improving spam detection activities in online social media.

Other works use textual features to improve their classifier, as those developed by Alsaleh et. al. [2] and Alarifi et. al. [1] that analyzed bot accounts to find the best set of features to be used in the classification step. They include: *i)* number of hashtags per tweet, *ii)* number of times a hashtag has been used, *iii)* number of links, *iv)* whether the profile picture contains a face, or it is the basic Twitter profile picture, *v)* mentions of different users with the same text, and *vi)* number of lists in which the user is listed on, are examples of selected features. With the features set, the authors used four machine learning algorithms for the classification step:

[1] Application Programming Interface

Decision Tree, Random Forest, Support Vector Machines (SVM), and Multilayer Neural Network.

Dickerson et. al. [7] proposed linguistic, network and application of variables to distinguish humans from bots using sentiment features. They presented SentiBot, a sentiment-aware architecture for the Twitter platform. On this architecture, they combined a set of features, such as sentiment features, neighborhood metrics, syntactic metrics, semantic linguistic models, graph-theoretic metrics, among others, collected from Twitter, related to the Indian elections in 2014. The users were manually labeled as bots or humans.

In these works, the main objective is to generate computational tool and methods to detect a bot user or a bot message leaving aside more information. However, they are not focused on assisting users in the task of detecting bots.

2.2 Media Literacy

Media literacy is an area that discusses the ability to access, analyze, evaluate and create messages in various contexts [12]. This area covers how the media is accessed (e.g., by television or internet), how the message is analyzed according to the reader's previous knowledge, how it is evaluated, and how a new message is sent forward.

This area has studies on how a media message need to be sent and how the receptor understands it. The user should be able to analyze and evaluate the media content, pondering the message relevance and confidence.

Fleming [8] conducted a case study at the Journalism School at Stony Brook University to implement a new form to present media literacy in the course to help students to assess the news quality. Through interview analysis, the news literacy form teaches students how to access, evaluate, analyze and appreciate journalism.

In [15], it is presented the main advantages of media literacy: 1) it promotes critical thinking skills to make independence choice, as which media select and how to interpret the information; 2) impacts individuals and society; 3) how to analyze and discuss a media message among others. Based on these advantages, we believe that media literacy is very important in the context of OSNs since this environment allows fast spreading of information. Such information is generated and consumed by its users, therefore they have responsibilities in how they handle with this information.

3 BUILDING THE CLASSIFICATION MODEL

In this section we present the dataset used in our experiment, the steps to prepare the data, and detail the classification process used to identify bots, as well as the method for obtaining the classification decision explanations.

3.1 Dataset

We chosen Twitter as OSN since it has a set of characteristics that favors this kind of research. An example is the hashtag (#), used as a subject mark in each message, so users can direct their messages to a specific theme. Another example is the mention (@), that allows to reply a user message or quote them.

One of the most important features that makes Twitter an useful OSN for study is its powerful API[2], which is well documented and allows a data sample for free download.

For our analysis we select a dataset labeled by Morstatter et. al. [14], that was collected from Twitter in the Arab Spring between 2011 and 2013, using the keywords: #libya, #gaddafi, #benghazi, #brega, #misrata, #nalut, #nafusa, #rhaibat, as well as a geographic bounding box around Libya. Using the Twitter API in 2015 they got account status from each user. Whoever was different from active was considered a bot. They provide a user list which includes the class (bot or human).

3.2 Preprocessing

We use the list of users labeled as users or humans to collect from Twitter all English messages from each user. An obstacle encountered was that some Twitter accounts no longer exist, which decreased the number of bot messages. This caused the dataset to be severely unbalanced towards human users.

To overcome this problem, we decided to select for each bot a group of human users for which messages most closely resemble the bots textual subjects. In this way, the comparison would be ideally done under textual cues that are less related to the topic, since the topic would be the same for both classes. The idea behind such alignment of topics is that by isolating the "topic feature" – a meta feature for the distribution of words in a message – the remaining textual features would be more easily captured by a Machine Learning classifier. On the other hand, such restriction of the word distribution also limits the ability to capture other possible textual features that are also unrelated to the topic. An alternative solution to this problem would be to separate the messages by topic and compare the bot messages with the human ones in all topics but the ones they share. Such solution was, however, not feasible under the time constraints for this study.

In summary, only users posting messages about subjects matching at least one bot were selected. For each bot, all human users were ranked by a relative word frequency score using the following formula:

$$score_{b_i,h_j} = \sum_{i,j} \sum_{k} F_{b_i}(w_k) * F_{h_j}(w_k) \quad | \; w_k \in W = W_{b_i} \cap W_{h_j} \tag{1}$$

where b_i is a bot user, h_j is a human user, $F_u(w)$ is the relative frequency of the word w in user's u messages, given by the quotient of the word count $\#w$ by the most frequent word count $\#w^*$. W_u is the word set of user u. Such score increases with the amount and frequency of shared vocabulary between two users.

The intuition behind the scoring formula is as follows:

- By calculating the relative frequency $F_u(w)$ of word w in user's u messages, a frequency vector $\hat{f}_u$ can be obtained, where each position refers to a single word in the shared vocabulary $W = W_{b_i} \cap W_{h_j}$.
- A measure of alignment between the user's word distributions can then be obtained by taking the dot product of

[2] Application Programming Interface

vectors $\hat{f}_u$, as it grows linearly with the relative frequency product $\hat{f}_{b_i} * \hat{f}_{h_i}$ of each word in the vocabulary.

- $score_{b_i,h_j}$ express the dot product $\langle \hat{f}_{b_i}, \hat{f}_{h_j} \rangle$ between frequency vectors for the bot user b_i and human user h_j, where $b_i \in B$, the set of bot users, and $h_j \in H$, the set of human users.

For each bot, the top n ranked human users were selected so that the number of messages were as close as possible. The selected users (bots and humans) messages constituted a balanced dataset, which was used for this work.

Before this step we had 9348030 messages in total, being 9036790 human messages and 311240 bot messages. The number of bot users is 106, while the number of humans before the preprocessing was 18963. After the preprocessing, we had 295307 human messages and 506 human users, almost 5 human users per bot and the number of messages between the classes balanced.

3.3 Analysis

To proceed with the analysis, we selected textual features to train a decision tree model classifier. Since one of our contributions is a user guide on how to detect a bot message, this approach was selected due to the ease of interpreting the resulting model. The decision tree algorithm implementation is *CART*[3]. The use of only textual features is a limiting factor regarding model accuracy, but provides way of classification that is reproducible by a human user with no further tools, given enough explanation about the path taken in the tree to reach a decision.

The selected textual features are shown in Table 1. We selected some features according the OSN characteristics, such as hashtag and mentions, that are common on Twitter. Part of Speech (POS) features were used to explain the message composition allowing superficial syntactic and semantic analysis. The selected POS were *noun, pronoun, verb, adverb, adjective, preposition, conjunction, numeral and interjection.*

Following the works of Alsaleh et. al. [2] and Alarifi et. al. [1], we used sentiment classification as a feature. The python[4] library *VADER* [9] was used to generate the values: positive, neutral and negative sentiment as 1, 0, -1, respectively, to our classifier.

4 EXPERIMENTS AND THE EXPLANATION METHOD

For evaluating model performance regarding the bot classification task, we implemented a 10-fold cross-validation, to train and test our classifier. To generate an intelligible tree and avoid overtraining, we set the tree maximum depth as 4 and the tree maximum leaf nodes as 7. Table 2 presents the mean and standard deviation from the training and test.

[3]https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier
[4]https://www.python.org/

Table 2: Model Results (10-fold CV)

Index	Mean	Std. Deviation
Accuracy	72%	0.002
Precision	68%	0.002
Recall	80%	0.003
F1	73%	0.002
Area under curve (AUC)	72%	0.002

With the decision tree simplification, some features showed to be more important than others. We found out that the POS-tag features, such as the number of interjection, conjunction, preposition, adverb, numeral and adjective present on the text does not contribute with the classification.
The same occurred for the number of hashtags, punctuation, and mentions. The sentiment classification is not an important feature of this dataset. The features used in the final classifier were the number of uppercase characters, the number of pronouns, the number of verbs and the number of links present in the message.

In the resulting tree, we observed that bot messages do not have much uppercase text (less than one), and the number of words categorized as noun and verb are less than five and one, respectively. As a recent update on Twitter [17], the number of allowed characters in the message was doubled (from 140 to 280), we could observe that messages with more than 140 characters have a considerable probability of being a bot message. Messages with less than 140 characters, but with less than one pronoun and one link in the text may also indicate a bot.

The next step was then describing the decision path in a human-interpretable way for a given message. For this, we traversed the decision path "translating" the attribute names to more understandable descriptions, and used a simple node separation syntax, as the following examples show:

Decision path (single): $qtd_upper < 1$ is **true** $\Rightarrow$ *bot*
Explanation: See Figure 1

Decision path (all human): $qtd_upper < 1$ is **false** $\rightarrow$
$qtd_text \leq 140$ is **true** $\rightarrow$
$pron_count \leq 1$ is **false** $\Rightarrow$ *human*
Explanation: See Figure 2

Figure 3 illustrates the full decision tree.

A simple qualitative evaluation was designed so that a decision path explanation would be able to be understood by at least 2 people: A function called $interpretability(expl(m))$ were $expl(m)$ is the explanation given by the tool on the decision path taken for the input message m. The function is $true$ for the case 2 people declared to understand the explanation, and $false$ otherwise. Due to the small size of the decision tree, we could cover all decision paths so that $interpretability(expl(m))$ was true for any path.

Table 1: Textual Features Selected

Feature	Description
Number of links	Number of URLs in the message
Number of mentions	Number of Mentions (@) in the message
Number of hashtags	Number of Hashtags (#) in the message
Number of punctuation	Number of punctuation in the message
Number of uppercase	Number of letters in uppercase in the message
Number of Nouns	Number of as nouns in the message
Number of Pronoun	Number of pronouns in the message
Number of Verbs	Number of verbs in the message
Number of Adverbs	Number of adverbs in the message
Number of Adjectives	Number of adjectives in the message
Number of Prepositions	Number of prepositions in the message
Number of Conjunctions	Number of conjunctions in the message
Number of Numeral	Number of numeral words in the message
Number of Interjections	Number of interjections in the message
Text length	Number of words in the message
Sentiment	Sentiment message classification

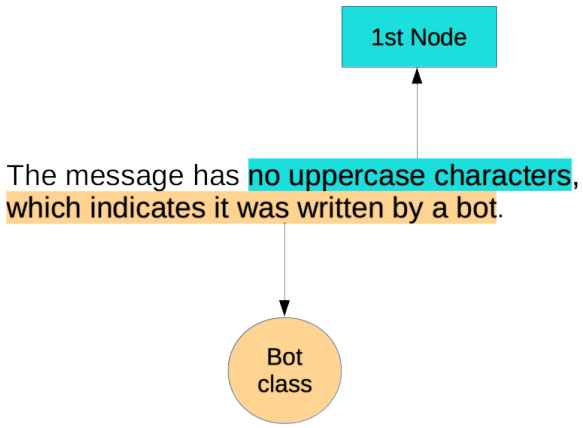


Figure 1: Explanation for the classification decision with decison path: *qtd_upper* < 1 is true ⇒ *bot*

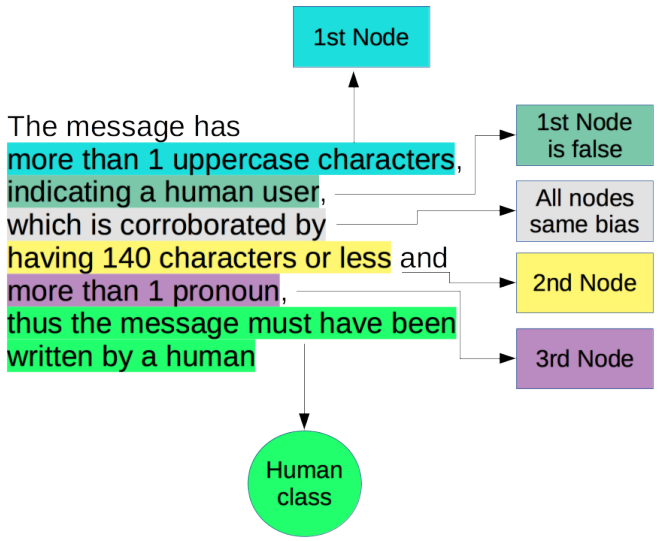


Figure 2: Explanation for the classification decision with decison path: *qtd_upper* < 1 is false → *qtd_text* ≤ 140 is true → *pron_count* ≤ 1 is false ⇒ *human*

The decision to keep the decision tree small was also made knowing that the bots are a moving target in terms of classification. As soon as the provided explanations become commonplace among the OSN users, bot creators will also update their systems to elude the classification criteria. However, by keeping the tree simple and so also the provided explanations, the users would focus on the most relevant cues for bot messages, making it easier for users to keep up to date with the changing criteria. This cat-and-mouse game presumes the frequent inclusion and exclusion of features on the classifier, which can capture the correct textual tracks left by the bot systems. Such features would inevitably go through Semantic Analysis, as bot systems become more sophisticated. The challenge then becomes keeping the explanation of such features (i.e., syntactic, semantic) simple, so the users can identify their instances to properly evaluate them.

5 USER INTERFACE

In this section, we present the user interface to the method developed in work. In order to facilitate use of the tool, we proposed a web browser plugin that allows the user to select the Twitter message and check if it may be a bot message and the explanation

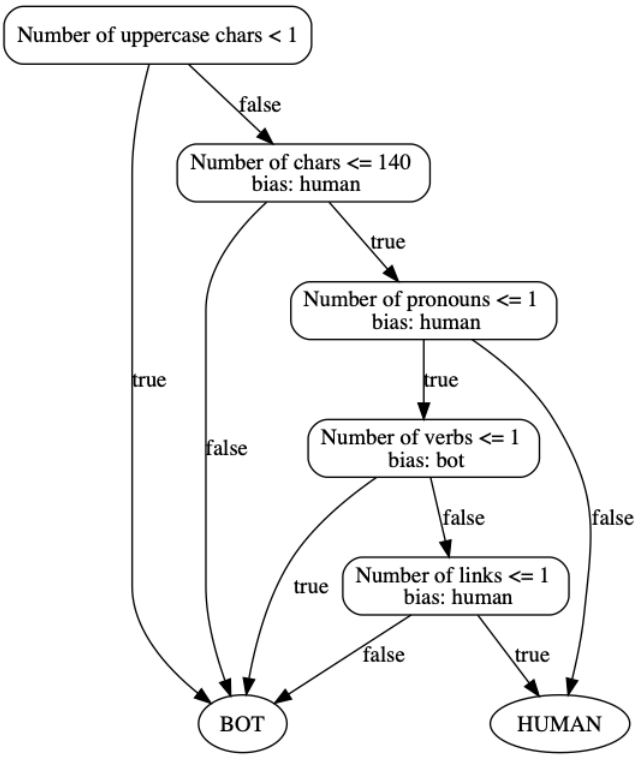


Figure 3: The complete decision tree. Bias values indicate the class that would be accepted if the decision path stopped at that node.

for the classification decision. This research step is intimately connected to the media literacy approach because it generates not only the message classification to the user but how this classification was achieved.

After the user installs the web browser plugin into his/her browser, it will be possible to select a message from the Twitter web site, and with a mouse left-click check if that message is a bot message.

The selection and classification process is shown on Figure 4. The web browser plugin sends the Twitter text to our API that accesses the trained model to check the message. The model returns the text classification with the explanation for the decision. The user will receive a communication that includes both returned information.

In Figure 5 we have an example of how the user will access the plugin. After selecting a message, with the mouse left-click it will be presented the menu with a "Check Message" option. After select this option, our API will receive the text and send to the model for classification. The return example is shown in figure 6, where the message is classified as written by a bot. In the case of a human message, the plugin returns the response shown in Figure 7.

In these figures, we used text from our datasets and explanation obtained from the decision tree classification path. In Figure 6, the message "@blabla[5] all of your tweets are answered in my blog postings." is classified correctly as a bot message, given that it has

[5]Username suppressed

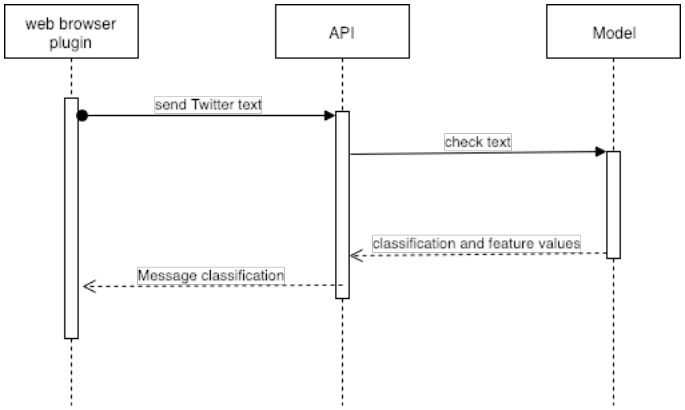


Figure 4: Sequence Diagram.

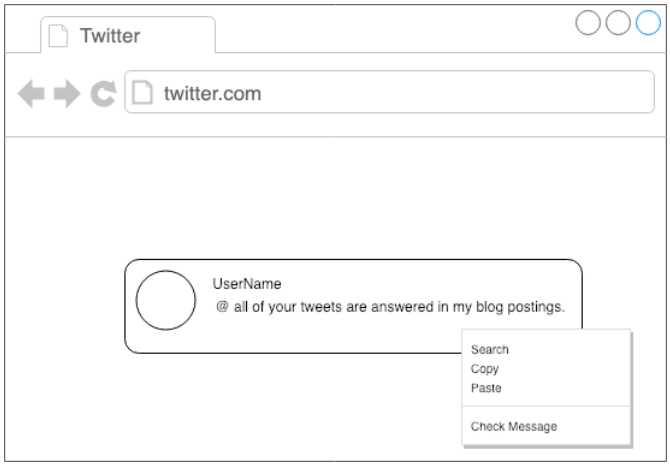


Figure 5: Plugin selection

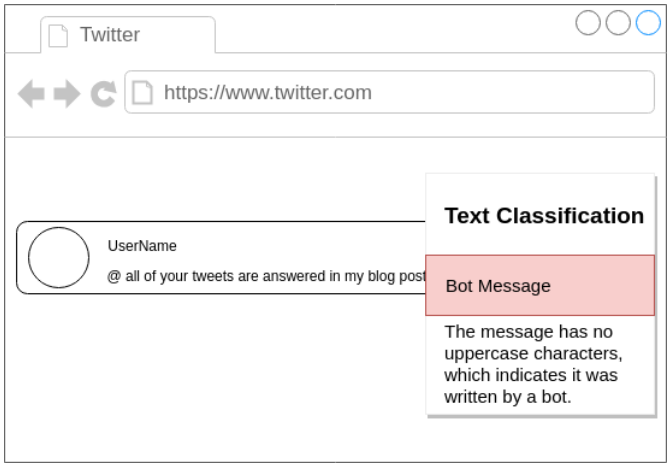


Figure 6: Plugin Response for Bot Message

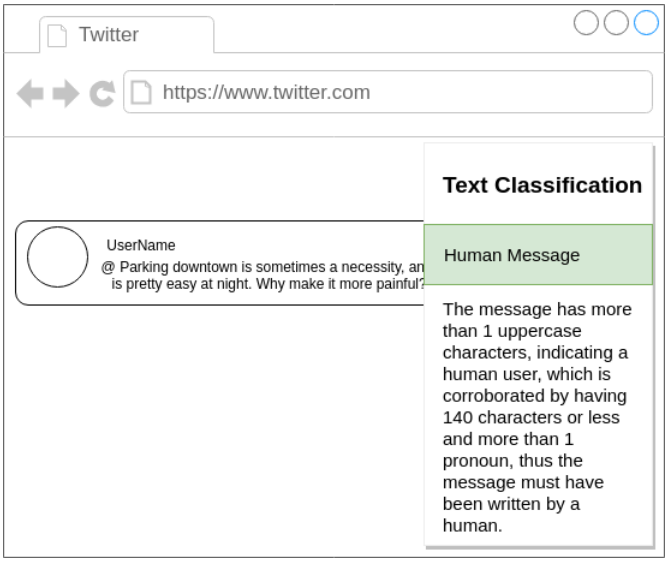


Figure 7: Plugin Response for Human Message

no uppercase letter, as explained in the previous section. It is an example of true positive test, where the model correctly detected a bot message.

The Figure 7, the message "@blabla[6] Parking downtown is sometimes a necessity, and usually is pretty easy at night. Why make it more painful? #SLC" is correctly classified as a human message, having 5 uppercase letters, a text length less than 140 characters and a pronoun. Based on our model, such characteristics are more related to the way that a human writes a message, rather then a bot. This case is an example of true negative test, where the classification is not a bot message (human message).

We observed in other example, the message "Supreme Court postponed to hear the appeal of Andualem Arage et al, who are jailed on terrorism charges, for November 22. #Ethiopia" has more than one uppercase letter (6), has less than 140 characters (131) and one pronoun. With this values, the model classify this message as a human message, however, it is a bot message. This classification is a false negative, where we have a wrong classification to a non-bot message.

A false positive example is a message that does not contain any uppercase letters, and is automatically classified as a bot message but instead belongs to a human user.

These misclassifications can be overcome through user feedback, which inform the tool about the mistake, so this can be used to improve our classifier. Additionally, collecting better textual features may improve the results, although those require further investigation.

Based on the main advantages introduced by [15], our proposed method shows how the message can be interpreted, given the message classification and how a user can look into media content before spreading a harmful messages that can have a serious impacts on Society.

[6]Username suppressed

6 CONCLUSIONS

This work presents a study that aimed to analyze textual features from bot and human written messages that were used in previous related studies, verifying the feasibility for agile bot detection while producing useful guidance for an ordinary user to recognize a bot message. This approach, based on media literacy, has the main objective to teach a user how one could detect an eventual bot message, which we believe contributes to non-dissemination of this kind of message.

We analyzed messages based on textual features of the dataset provided by Morstatter et. al. [14], composed by Twitter messages from the Arab Spring event and labeled as bot and human users. The use of only text is the simplest way for an ordinary user to be aware that there may be something wrong with the message. The combination of many characteristics, such as the number of followers, number of friends, among others, is much more difficult for a human to check the many possibilities and judge if the text was written automatically.

The decision tree machine learning algorithm was used to select the best features and generate the explanation to teach the user how to recognize a bot message. We simplify the tree structure to reach an intelligible model, at a minor cost on model accuracy.

A web browser plugin is proposed, that will help the user check if a message is a bot message or not. This plugin sends the text to our trained model through a service API, which in turn returns to the plugin user the message classification and an explanation on the steps to achieve the result. In this way, users will check the message classification and also learn how to classify messages on their own.

The method proposed in this paper is different from related works due to the focus being not only to bot and human messages using machine learning, but use the obtained model to teach an OSN user how to distinguish bots from humans.

As a limitation, the model accuracy is considerably under the current state of the art for Twitter data and need to be improved. The use of textual features only decreased the accuracy in exchange for simplicity. A way of improving this would be to include alternative textual features, such as event and sub event relation, among others. Such forward solutions should also cover the constant evolution of bot systems, which will adapt to the criteria exposed by our system, while keeping the users up-to-date with the ever changing criteria.

For the next works we will improve the model, adding more textual features, without losing interpretability. We will also apply this approach in other datasets, we will compare messages along time, subject or events and also messages in different languages. In this way, we can examine how such textual patterns are build and changed. Conducting acceptability tests with real users based on the web browser plugin will be a future task.

REFERENCES

[1] Abdulrahman Alarifi, Mansour Alsaleh, and Abdul Malik Al-Salman. 2016. Twitter turing test: Identifying social machines. *Information Sciences* 372 (dec 2016), 332–346. https://doi.org/10.1016/j.ins.2016.08.036

[2] Mansour Alsaleh, Abdulrahman Alarifi, Abdul Malik Al-Salman, Mohammed Alfayez, and Abdulmajeed Almuhaysin. 2014. TSD: Detecting sybil accounts in twitter. In *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*. IEEE, 463–469. https://doi.org/10.1109/ICMLA.2014.81

[3] S. Aslan. 2018. Twitter by the Numbers: Stats, Demographics & Fun Facts. https://www.omnicoreagency.com/twitter-statistics/

[4] Sajid Yousuf Bhat and Muhammad Abulaish. 2013. Community-based features for identifying spammers in online social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*. ACM Press, 100–107. https://doi.org/10.1145/2492517.2492567

[5] Aaron Black. 2018. A new study suggests fake news might have won Donald Trump the 2016 election. https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/?noredirect=on&utm_term=.1029a721270f

[6] P. Cooper. 2019. 28 Twitter Statistics All Marketers Need to Know in 2019. https://blog.hootsuite.com/twitter-statistics/

[7] John P. Dickerson, Vadim Kagan, and V. S. Subrahmanian. 2014. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?. In *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 620–627. https://doi.org/10.1109/ASONAM.2014.6921650

[8] Jennifer Fleming. 2014. Media literacy, news literacy, or news appreciation? A case study of the news literacy program at Stony Brook University. *Journalism and Mass Communication Educator* 69, 2 (jun 2014), 146–165. https://doi.org/10.1177/1077695813517885

[9] C.J. Hutto and E.E. Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)* (jun 2014).

[10] Rodrigo Augusto Igawa, Sylvio Barbon, Katia Cristina Silva Paulo, Guilherme Sakaji Kido, Rodrigo Capobianco Guido, Mario Lemes Proenca Junior, and Ivan Nunes da Silva. 2016. Account classification in online social networks with LBCA and wavelets. *Information Sciences* 332 (mar 2016), 72–83. https://doi.org/10.1016/j.ins.2015.10.039

[11] Bernardo Pereira Lauand. 2016. *Abordagem Experimental para a Localizacao e Detecao de Eventos em Tweets*. Master's thesis.

[12] Sonia Livingstone. 2004. Media literacy and the challenge of new information and communication technologies. *Communication Review* (2004).

[13] Juan Martinez-Romo and Lourdes Araujo. 2013. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications* 40, 8 (jun 2013), 2992–3000. https://doi.org/10.1016/j.eswa.2012.12.015

[14] Fred Morstatter, Liang Wu, Tahora H. Nazer, Kathleen M. Carley, and Huan Liu. 2016. A new approach to bot detection: Striking the balance between precision and recall. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. IEEE, 533–540. https://doi.org/10.1109/ASONAM.2016.7752287

[15] Art Silverblatt, Andrew Smith, Don Miller, Julie Smith, and Nikole Brown. 2014. *Media Literacy: Keys to Interpreting Media Messages* (4 ed.). 549 pages.

[16] J. Sommerland. 2018. Russian Hackers Used Tumblr To Spread 'fake News' During Us Election, Company Reveals. https://www.independent.co.uk/life-style/gadgets-and-tech/news/tumblr-russian-hacking-us-presidential-election-fake-news-internet-research-agency-propaganda-html

[17] A. Sulleyman. 2017. TWITTER INTRODUCES 280 CHARACTERS TO ALL USERS. https://www.independent.co.uk/life-style/gadgets-and-tech/news/twitter-280-characters-tweets-start-when-get-latest-a8042716.html

[18] J. Swaine. 2018. Twitter admits far more Russian bots posted on election than it had disclosed. https://www.theguardian.com/technology/2018/jan/19/twitter-admits-far-more-russian-bots-posted-on-election-than-it-had-disclosed

Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists

Piotr Sapiezynski
Northeastern University
p.sapiezynski@northeastern.edu

Wesley Zeng
Northeastern University
zeng.wes@husky.neu.edu

Ronald E. Robertson
Northeastern University
robertson.ron@husky.neu.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

In this work, we introduce a novel metric for auditing group fairness in ranked lists. Our approach offers two benefits compared to the state of the art. First, we offer a blueprint for modeling of user attention. Rather than assuming a logarithmic loss in importance as a function of the rank, we can account for varying user behaviors through parametrization. For example, we expect a user to see more items during a viewing of a social media feed than when they inspect the results list of a single web search query. Second, we allow non-binary protected attributes to enable investigating inherently continuous attributes (e.g., political alignment on the liberal to conservative spectrum) as well as to facilitate measurements across aggregated sets of search results, rather than separately for each result list. By combining these two elements into our metric, we are able to better address the human factors inherent in this problem. We measure the whole sociotechnical system, consisting of a ranking algorithm and individuals using it, instead of exclusively focusing on the ranking algorithm. Finally, we use our metric to perform three simulated fairness audits. We show that determining fairness of a ranked output necessitates knowledge (or a model) of the end-users of the particular service. Depending on their attention distribution function, a fixed ranking of results can appear biased both in favor and against a protected group[1].

CCS CONCEPTS

• **Information systems** → **Page and site ranking**; **Content ranking**; • **Human-centered computing** → *User interface design*.

KEYWORDS

information retrieval; group fairness; ranked lists

ACM Reference Format:
Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317595

[1]The repository containing Python code associated with this work can be found at https://github.com/sapiezynski/fairness-attention

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317595

1 INTRODUCTION

The exponential growth of information available online has necessitated the development of Information Retrieval (IR) algorithms that decide what content is *relevant* to users. For example, upwards of 5.5 billion searches are conducted on Google every day [18], and in response to each, Google filters and sorts a list of ~15 results [46]. Similarly, over 1.4 billion people visit Facebook daily [25] and scroll through a list of content from friends and advertisers deemed relevant by the News Feed algorithm. Finally, tens of millions of worker profiles are available on LinkedIn, filtered and sorted when recruiters search for prospective employees.

Recently, a concern has been growing that even seemingly benign IR systems may negatively impact people. It has been shown that algorithms can reflect societal biases [2], and ranking mechanisms are no different. Kay et al. found that Google Image Search returned images portraying men and women in stereotypical roles in response to occupation-related queries, and that these results reinforced stereotypical gender roles [34]. Others have examined partisan slant in search results [12, 38, 46] in light of user studies demonstrating that partisan search results can significantly influence voting behavior [15, 16]. Lastly, two studies have examined the relationships between gender, race, and ranking of job seekers on employment websites [9, 23], where systematic biases that push members of protected classes into lower ranks could result in the loss of employment opportunities and earnings [36].

Only recently have researchers started addressing the problem of fairness in ranked outputs. From the IR side, this includes novel ranking algorithms that aim to achieve *representational parity* (a.k.a. *group fairness*): the ranker is required to assign a certain fraction of top ranks to people in the protected or minority class) [3, 8, 50, 54]. From the *algorithm auditing* side (i.e., investigators who look for fairness problems in black-box systems) [48], Yang and Stoyanovich introduced metrics for quantifying whether the outputs from a given search engine are group-fair [51]. This enables auditors to examine real-world search engines and hopefully hold them accountable for producing unfair outputs.

However, there is a reoccurring challenge in the extant literature on fair ranking: how to model user attention? Eye-tracking studies and click-stream data show that users do not distribute their attention evenly over ranked lists of information [13, 19, 21, 42].

This unequal distribution of attention must be taken into account when designing fair IR systems and evaluating whether a given IR system is fair. The trouble is that the distribution of attention for a given search engine may be unknown, since it varies based on the user interface of the service (e.g., pagination boundaries) and context (e.g., searching for a specific movie trailer versus searching for a new employee).

Most of the previous work on fairness in ranked lists has assumed logarithmic discounting of attention [50, 51]. However, because of its flattening shape for low ranks, logarithmic drop-off is impractical for modeling attention. For example, modeling attention this way would mean assuming that on a list of 100 results, the sum of attention given to last eight results is bigger than the attention paid to the first. Biega et al. use a geometric distribution instead, but do not investigate the consequences of varying its steepness [3].

In this work, we extend the literature on fair ranking by introducing a novel metric for measuring group fairness in ranked outputs. Our metric, the *Viable-*Λ *Test*, is designed for auditors and answers two questions: (1) does there exist a distribution of user attention $P(\Lambda)$ such that the output of a search engine is group fair, and (2) if so, what is the parameterization $\Lambda = (\lambda_1, ..., \lambda_m)$ of this distribution? In contrast to prior work that attempts to "score" the fairness of a ranking algorithm [51], our metric fundamentally re-frames the question of fairness to involve the consumer of the ranking and their attention. If the fitted model of attention $P(\Lambda)$ does not match empirical observations of user attention in the given search engine, then the system does not achieve representational parity.

Overall, our paper makes three key contributions:

(1) We introduce a novel metric, the Viable-Λ Test that binds the usage patterns of a list to the measurement of fairness.
(2) We enable fairness measurements in situations with class assignment uncertainty, results aggregation, multiple protected classes, and continuous protected variables.
(3) We evaluate the Viable-Λ Test on data from three real-world services: a resume search engine, a dating service, and a web search engine. Our results demonstrate that the choice and parameterization of the attention function can lead to dramatically different conclusions about whether (and how) the rankings are biased.

Note that it is not possible to determine with certainty whether a given set of search results are biased without knowing the true attention distribution function for users of the corresponding service. Therefore, our work should *not* be seen as an audit study, but rather a showcasing of a metric that can be used by the operators or internal auditors of these services to ensure fair delivery of results.

The remainder of the paper is organized as follows. In Section 2 we introduce related work on fairness in ranked lists and auditing of ranking algorithms. In Section 3 we provide a detailed description of the mechanics of the proposed metric. In Section 4we explain how the parameters of our metric should be set and interpreted based on context of its use. In Section 5 we analyze three case studies: a hiring service, a dating service, and Google search. In Section 6 we explain the limitations of our approach. Section 7 suggests the directions for further research, and Section 8 concludes the article.

2 BACKGROUND

In Pursuit of Fairness. As use of large, observational datasets has proliferated, so have concerns that systems leveraging this data may have a negative impact on people. The machine learning community has mapped the legal concepts of *disparate treatment* and *disparate impact* to *direct* and *indirect discrimination* by algorithms, respectively [5, 22, 43]. Zafar et al. introduced the concept of *disparate mistreatment* to refer to situations where false positives and negatives are not equally distributed across subpopulations [52, 53]. Such situations have been shown to occur e.g., in pre-trial assessments [1] and academic performance predictors [49].

While direct discrimination can be corrected by removing the protected attributes from the data, indirect discrimination is more challenging to address. Dwork et al. proposed two potential objectives for mitigating indirect discrimination: under *individual fairness*, similar people should be treated similarly by the algorithm, while under *group fairness* subpopulations should be treated equivalently to the whole population [14]. There is a large and growing literature on how to achieve these objectives in machine learning-based classifiers [5–7, 14, 17, 22, 28–33, 39, 55].

Fairness in Ranking. Achieving fairness in IR systems has received less attention in the academic community. One challenge is that research on fair classification does not necessarily generalize to the ranking context. A second challenge is accounting for *order effects* [41], i.e., the well-established tendency of human beings to pay more attention to items at the top[2] of a ranked list.

A few methods for generating group-fair search results have been proposed. Zehlike et al. leverage randomization by positing that a given ranked output is fair if it could have been generated by a random Bernoulli process [54]. Celis et al. propose a more general approach allowing the user to specify the fairness constraints [8]. Unfortunately, neither of these take user attention into account: their methods do not distinguish between different orderings of a set as long as a minimum fraction of items from the minority class are presented at each rank. In contrast, Singh and Joachims argue that even if the ranking itself is unbiased, small differences in placement may lead to large discrepancies of attention [50]. Biega et al. point out that any single ranking of similarly relevant items is individually unfair because of the uneven distribution of attention [3]. Therefore, they propose achieving individual equity (attention corresponding to relevance) within a certain number of realizations by systematic reshuffling of the list.

Algorithm-in-the loop Approach. Most of the work we discussed so far focus on measuring or correcting the algorithm without explicitly involving its users. In contrast, Green and Chen emphasize the need for considering the whole sociotechnical system [20]; they show that rather than focusing solely on the bias in an automated risk-assessment system, one needs to include contextual information on the system is actually used by judges and how it affects their decisions.

Auditing Search Engines. There is a growing body of work from the algorithm auditing [48] community that aims to measure

[2]We use "top" and "high" to refer to the numerically lowest ranks in lists, e.g., rank one, in keeping with the norms of the IR literature [10, 27].

whether real-world search engines are fair and unbiased. Kay et al. found that Google Image Search presented results that were stereotypically gendered [34], while Hannak et al. and Chen et al. showed that search engines on employment websites were not group fair with respect to race and gender [9, 23]. Audit studies have also examined the political partisanship of search results from Twitter and Google [12, 38, 46].

An open challenge for the auditing community is selecting appropriate metrics for assessing whether search engine results are group fair. For example, Kay et al. only looked at simple metrics like average representation that fail to take order effects into account [34]. Other audits have used group representation in top K ranks [24], logarithmic discounting [9, 23, 51] and linear normalization by rank [38, 46] to model the decay of attention. In this work, we argue that these ad hoc methods do not accurately model users' attention, and may lead to incorrect conclusions about (un)fairness of IR systems.

3 METHODS

In this section, we introduce a novel metric for measuring group fairness in search results. This metric, the Viable-Λ test, combines existing research from algorithmic auditing, IR, and human-computer interaction to address the human factors inherent in this problem.

3.1 Overview

Suppose we are given an ordered list of search results $R = [r_1, \ldots, r_n]^T$. Our goal is to measure the representation of some target property p that is shared by each $r_i \in R$. The metric proposed in this paper requires the auditor to specify the following five elements: (1) an *alignment vector* L_R, (2) an *attentional weight vector* W_R, (3) a *population estimator* $\hat{p}$, (4) a *distance metric* d, and (5) a *maximum allowable distance* $\delta_{\max}$.

Below we briefly introduce these elements and explain their role in the Viable-Λ Test. In Section 4 we explain the design choices behind each element and how to allocate them appropriately in practice.

The **alignment vector** L_R is a vector of probability distributions $[l_1, \ldots, l_n]^T$ that describe the group membership (or *alignment*) of r_i with respect to the target property p. The subscript $_R$ indicates that L_R has the same length as R, and that each $l_i \in L_R$ corresponds to the alignment of each respective $r_i \in R$. l_i can be either discrete or continuous.

The **attentional weight vector** W_R is a probability vector $[\omega(r_1), \ldots, \omega(r_n)]^T$ that models the relative user attention allocated to each $r_i \in R$. While it is difficult to determine the exact distribution of W_R, we can make assumptions about its *shape*. Formally, $W_R \sim P(\Lambda)$ where P is a family of n-truncated discrete probability distributions with an unknown set of true parameters Λ.

We calculate the expected cumulative exposure E_R of group representation in R by taking the dot product of L_R and W_R:

$$E_R = L_R^T \cdot W_R \tag{1}$$

Note that E_R is a probability distribution with the same domain as each $l_i \in L_R$ (i.e., a distribution describing the target property p).

The **population estimator** $\hat{p}$ is a probability distribution that estimates the true demographics of p. For reasons described in § 4.3, the following formula is often a well-motivated choice:

$$\hat{p} \overset{\text{def}}{=} \bar{L_R} = \sum_{r_i \in R} \frac{l(r_i)}{n} \tag{2}$$

where $\hat{p}$ has the same domain as l_i.

The **distance metric** d quantifies the statistical difference between the probability distributions E_R and $\hat{p}$.

The **maximum acceptable distance** $\delta_{\max}$ is the threshold that separates group fair from unfair search results. It is chosen in a context-dependent manner in conjunction with d. As we describe in § 4.4, d is essentially a statistical significance test, and δ_{min} is the test statistic threshold. Both components draw inspiration from the principles of traditional sampling statistics.

3.2 The Viable-Λ Test for Representational Parity

Given $L_R, \hat{p}, d$, and $\delta_{\max}$, we define group exposure as $E_R = L_R^T \cdot W_R$. Assume that $W_R \sim P(\lambda_1, \ldots, \lambda_m)$ where P is an n-binned discrete probability distribution with an unknown set of true parameters $\Lambda = (\lambda_1, ..., \lambda_m)$ in some domain space D with known bounds. Then R is unfair if:

(1) $\nexists \lambda \in D$ such that $d(e_r, \hat{p}) < \delta_{\max}$, i.e., there is no way to parameterize the attention distribution such that representational parity is attained; and
(2) For Λ satisfying the above condition, $W(\Lambda)$ matches reasonable expectations and data about true user behavior.

4 DESIGN CHOICES

This section describes the key components of Viable-Λ and the motivating decisions behind their design.

4.1 The Alignment Vector L_R

Alignment as a Probability Distribution. Given an ordered list of result items $R = [r_1, \ldots, r_n]^T$, we use an *alignment function* ι to map each result r_i to a probability distribution describing its *alignment* in terms of the target property p (e.g., race, gender, political alignment, etc.). As a motivating example, consider a resume search engine [9] that returns a list of job candidates R in response to a query (e.g., "software engineer"). Suppose we want to model the gender alignment p of these results; namely, p is a discrete probability distribution across three classes: 'Male,' 'Female,' and 'Unknown.' If the search results *explicitly* display the gender of each candidate r_i, then we can define alignment as follows:

$$l_i \overset{\text{def}}{=} \begin{cases} \{\text{Female: 1.0, Male: 0.0, Unknown: 0.0}\} & \text{if } r_i \text{ is female} \\ \{\text{Female: 0.0, Male: 1.0, Unknown: 0.0}\} & \text{if } r_i \text{ is male} \\ \{\text{Female: 0.0, Male: 0.0, Unknown: 1.0}\} & \text{otherwise} \end{cases} \tag{3}$$

In realistic scenarios, defining l_i may not be so trivial. Continuing our example, resume search engines typically do not explicitly state the gender or race of job candidates. However, recruiters (and auditors) may still be able to infer them using other information, like

a user's profile picture, given name, etc. Most existing approaches to measuring fairness assume that all l_i are explicitly known and binary (canonically between protected and non-protected classes S and S^C) [14, 51]. However, this assumption has not generalized to empirical studies. By using a probability distribution rather than a binary indicator, we cover the cases where the class assignment is ambiguous, where there are multiple classes, or there are multiple realizations of the ranking.

4.2 Attentional Weights W_R

W_R encapsulates the well-documented fact that search engine users do not treat all search results equally [10, 19, 21, 45]. For example, the first result on Google Search is estimated to receive approximately 30% of all clicks, and the results on the first page account for approximately 90% of clicks [35]. This observation remains true to a large extent even if the order of the search results is inverted [35].

Modelling Attentional Weights ω. W_R is a probability vector that models the amount of user attention that each $r_i \in R$ receives. $W_R = [w_1, \ldots, w_n]^T$, where $w_i = \omega(r_i) \in [0, 1]$ is the output of a weighting function ω applied to r_i. ω is influenced by (1) the user interface design of the search engine (for example pagination or highlighted results) and (2) the use context of the search engine. For an intuition on the latter, consider that a user may only view several web search results before finding an acceptable answer to their query, whereas that same user might view dozens of resumes from a resume search engine if they are tasked with hiring a new employee.

Fitting W_R from Empirical Data. A tempting way to approximate W_R is to use empirical data such as organic Click Volume (CV) and Click-Through Rate (CTR), widely used in digital marketing and by search engine proprietors. Unfortunately, there are three serious impediments to using an empirically derived W_R. First, click data is often proprietary, and thus unavailable to an external auditor. Second, click data is only an approximation for what users see in search results. Truly measuring user attention might require expensive eye tracking studies [10, 19, 21, 45]. Third, as we noted above, the way users distribute their attention over search results depends on website design and use context. Thus, although eye-tracking studies are available for sites like Google Search [10, 19, 21, 45], the data may not be applicable to other search engines.

Potential Choices for W_R. In algorithmic auditing, we need a way to measure user attention without (1) having access to the vendor's analytic data, and (2) having to conduct multiple eye tracking studies. We partially address this issue by making assumptions about the shape of W_R. Formally, we assume that $W_R \sim P(\Lambda)$ for some discrete truncated probability distribution P. W_R should meet two criteria:

(1) W_R is an n-truncated discrete probability distribution.
(2) Higher-ranked results receive substantially more attention than lower-ranked ones; i.e., for reasonably large n, $\omega(r_1) \gg \omega(r_i)$ as $i \to n$.

Some families of distributions that fit the above criteria are presented in Figure 1 and include Truncated Geometric Distribution, Truncated Log-series Distribution, and Truncated Discrete Pareto distribution [37]. For the remainder of the paper we use the truncated geometric distribution but in an actual measurement scenario, another choice may be more appropriate.

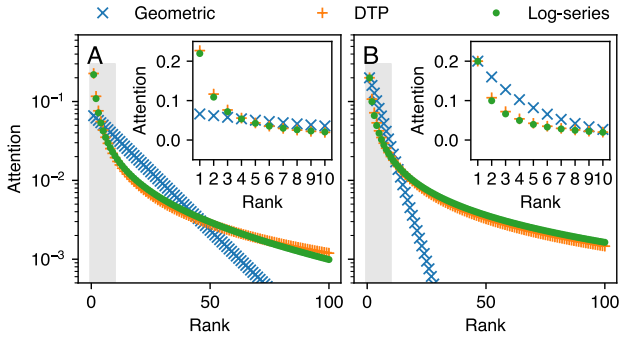


Figure 1: Comparison of Geometric, Log-series, and Discrete Truncated Pareto (TDP) on a list with length $n = 100$. In (A), the parameters are set so that $E[W_R]$ (i.e., the mean number of results seen) is 15. In (B), the parameters are chosen so that $\omega(r_1) = 0.2$.

The Case of Small n. The above distributions are applicable when n is reasonably large, i.e., well beyond the human attention span. When n is small, we can draw inspiration from psychology: Miller's Law famously states that the average human's working memory can hold roughly 7 objects at a time [40]. When n is sufficiently small, we expect the user to read all of the results.[3] Formally, when n is small, $W_R \tilde{} unif\{0, n\}$. Note that when choosing $W_R \overset{\text{def}}{=} unif\{0, n\}$ and $\hat{p} \overset{\text{def}}{=} L_R$, E_R is precisely equal to $\hat{p}$; therefore all short lists exhibit approximate representational parity.

The Problems of Inverse Log Scaling. We notably left out inverse log:

$$W_R = normalized\left(\left[\frac{1}{\ln(2)}, \ldots, \frac{1}{\ln(n+1)}\right]^T\right). \quad (4)$$

Inverse logarithmic scaling is commonly used in IR relevancy metrics such as nDCG, and also appears in some metrics of fairness in ranking [51]. However, this choice of W_R has two major flaws: *first*, it decays at a very slow rate and does not meet the relative convergence requirements described above in criteria (2). Even when $n = 1000$, $\frac{\omega(r_1)}{\omega(r_{1000})}$ is approximately 9.964, which implies that the last 10 search results are, in aggregate, as influential as the first result. *Second*, since there are no parameters for this choice of W_R, it incorrectly assumes that user behavior is the static across all platforms and for all search queries.[4]

4.3 Estimating Population Demographics $\hat{p}$

Our metric requires the specification of an estimator $\hat{p}$. This section describes how to choose $\hat{p}$ so that it serves as a reasonable prior for statistical parity.

[3] This may occur in practice when a user queries for obscure information or the vendor lacks data relevant to the search query
[4] The base of the logarithm is irrelevant; after normalization, they all evaluate to the same vector of values.

Implicit Estimators for p. A tempting proposal would be to choose an *implicit estimator* for p based on intuition or observational data. This is plausible when the search query Q is relatively simple. For example, consider the case when the user is querying a resume search engine for a list of certified nurses: recent data shows that the US national gender ratio for this profession is approximately 9.5:1 female to male [44], thus we expect fair rankings for this query to reflect this. For more complex queries, however, demographic data may unavailable. If the user instead queries for Android Developers in Greenville, AL with at least three years of experience, we lack an external data reference. Thus, we are unable to choose nor justify an implicit estimator $\hat{p}$.

Furthermore, Q itself cannot be directly examined in many IR systems. Online services rank their feeds using proprietary algorithms that rely on personalization. From an auditor's standpoint, a generalized fairness metric must have the ability to estimate p regardless of whether or not we have access to Q.

Choosing $\hat{p} \stackrel{\text{def}}{=} \bar{L_R}$. An IR-motivated alternative to circumvent these issues is to determine $\hat{p}$ based on the vendor's data R. Suppose that within an IR system, the vendor evaluates Q then filters their corpus to yield a subset of results R in which all $r_i \in R$ meet some relevancy threshold. Then we can calculate $\hat{p}$ as defined in Equation 2, i.e., an equally-weighted sum $\forall l(r_i) \in L_R$. Since this calculation relies on the vendor's data R, the validity of $\hat{p}$ is dependent on the integrity of the vendor's data. Thus, it is imperative to first audit the vendor's *data curation* for sampling bias and *result scoring* for direct discrimination. A hypothetical example of unfair data curation would be a job site that refuses to add female software engineering candidates to their database. Unfair query evaluation could occur when a vendor has female candidates in their database but fails to show them when a recruiter queries for "software engineers" (i.e., being a woman directly impacts the relevancy score). Fortunately, this style of audit is often straightforward: Chen et al. tested for direct discrimination in scoring by posting two identical resumes that varied only by gender, and showing that they appeared at directly adjacent positions in search results [9]. If a preliminary audit finds that any of these assumptions are jeopardized, we can immediately deem R as unfair without needing to calculate $\hat{p}$ or other components of Viable-Λ.

Consequences. In choosing $\hat{p} \stackrel{\text{def}}{=} \bar{L_R}$, our prior for statistical parity is fitted to the vendor's data. Thus, if the vendor's knowledge of p is lacking or biased (for example, unknowingly exhibits sampling bias), our estimate of p will be as well. In this situation, Viable-Λ becomes a metric of how well the vendor's ranking *represents their own knowledge of p*.

4.4 Distance Metric d and the $\delta_{\max}$ Threshold

We use d and $\delta_{\max}$ to delineate an *acceptance region* around $\hat{p}$.

Distance Metric d. d is a statistical distance metric that quantifies the difference between the two probability distributions E_R and $\hat{p}$. The choice of d follows naturally based on the domain of E_R and $\hat{p}$. In Section § 5, we demonstrate the use of Z-approximation for the binomial test statistic when E_R and $\hat{p}$ are both binomial distributions.

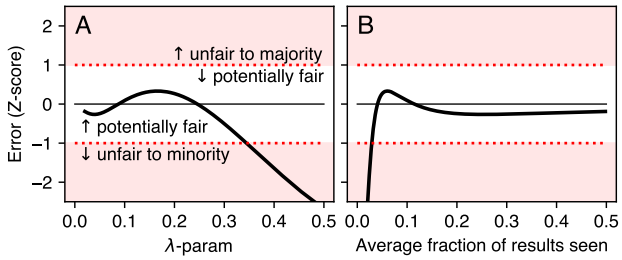


Figure 2: We find the range of the λ-parameter (and, thus, the corresponding average fraction of seen results (B)) for which the result list is potentially fair. If the attention given to a group is more than $1Z$ away from its representation in the population, we deem the results unfair.

Maximum Acceptable Distance $\delta_{\max}$. $\delta_{\max}$ establishes an acceptable range of values around $\hat{p}$ in which we can safely assume that group fairness is preserved. In general, $\delta_{\max}$ is the *test statistic* threshold to the statistical significance test d. In conjunction, d and δ_{min} constitute a statistical significance test with h_0: $E_R \sim \hat{p}$, i.e., E_R has the same sampling distribution as $\hat{p}$. A well-constructed $\delta_{\max}$ takes the following factors into consideration:

(1) *Statistically significant difference between E_R and $\hat{p}$ (size of R).* Suppose that we perform a search to yield results R with length $n = 100$, for some binomial alignment property. One interpretation of this scenario is that our result R is a sampling distribution of size $n = 100$ from a universe U of *all* relevant candidates. We can formalize this by saying that R is a simple random sample of U, where $n = 100 \ll |U|$.[5] U has a true binomial parameter p. Thus, we can employ statistical inference to construct a confidence interval for p. For a binomial distribution, the maximum likelihood estimate (MLE) of p yields $\hat{p} = \bar{L_R}$ with standard error $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{100}}$. By specifying a confidence level (e.g., 95% corresponding to $z = 1.96$), we can build a confidence interval for p. Thus, we cannot reject H_0 if E_R falls inside this range of values. For categorical p, we can check for statistically significant difference between $\hat{p}$ and E_R using the appropriate test statistics (e.g., Pearson's χ^2 Test for Independence).

(2) *Number of Search Queries Made.* This factor is applicable only in the case where the the vendor randomizes their search results. We can model $l(r_i)$ as the sampling distribution of S across k search realizations. We calculate the standard error in the same manner as above using k in lieu of n. An important distinction is that the previous standard error calculated with n represents uncertainty in p, while this standard error calculated with k represents uncertainty at each rank in $l(r_i)$. In the case where the vendor does not randomize their rankings or when the auditor can make an unlimited number of queries, this factor is irrelevant.[6]

[5] Frequentist statistics relies on the assumption that $n \ll |U|$, which reflects the fact that it is often impractical or expensive to collect a census over the entirety of U.

[6] With an unlimited number of queries, we can arbitrarily reduce the standard error to any $\epsilon > 0$. This is not possible with the confidence interval described in the previous section since it is the vendor—not the auditor—who determines the size of R. In the case where the vendor is able to increase n (i.e., by acquiring more data), then both the auditor and the vendor will be able to make more accurate estimates of p.

Plotting $d(E_R, \hat{p})$. In the case where W_R has one parameter, we can plot $d(E_R, \hat{p})$ as a function of λ. An example of this is illustrated in Figure 2.

4.5 Restricting the Domain of the Parameter Space D

In our model, we accept the null hypothesis that a ranking R is fair if there exists a set of parameters Λ within the parameter space D which brings us within an acceptable range of parity. However, we are susceptible to type-II errors since not all $\Lambda \in D$ match realistic expectations about user behavior. Thus, by truncating our parameter space D, we can increase the *power* (the probability of *not* making a type-II error) of our hypothesis test.

Context-based Assumptions. To illustrate this, consider a ranked list with length $n = 100$. Suppose that we have deduced that $W_R \sim Geom(\lambda, 100)$. The mathematically-permitted domain of the success parameter is $\lambda \in (0, 1)$. However, if we know for certain that $E[W_R]$ is between 2 and 50 (i.e., the average user views between 2 and 50 profiles), thus we can restrict our domain space to $\lambda \in (0.02, 0.5)$.[7]

Empirically-informed Truncations. Furthermore, we can also set up a small-scale experiment to estimate $E[W_R]$. Suppose that we have a small group of users with $N = 16$. Assume that the average number of results viewed is approximately normal with $\bar{x} = 27$ and standard error $s = 12$. Using the maximum likelihood estimator (MLE) we can construct a 95% confidence interval for $\bar{x}$ between approximately 21 and 33. Since $E[W_R] \approx 1/\lambda$, we can say that the parameter λ lies between 1/21 and 1/33, corresponding to the interval $(0.03, 0.47)$[8].

As we can see in the above example, it is possible to leverage data from small-scale experiments to estimate the likelihood of λ. This offers us an avenue to improve the statistical power of the Viable-Λ test.[9]

4.6 Generating Fair Ranked Lists

Other researchers have focused on creating fair ranked lists and their cost in terms of individual fairness in depth [3, 8, 50, 54]. Here, we provide a few examples of ranked lists that are fair for fixed λ. Figure 3 shows four best-attempt fair lists with varying class imbalance (1:10 in top row, 5:10 in bottom row; minority class A with light blue, majority class B with dark blue) and attention distributions ($\lambda = 0.1, 0.5$ from left column to right). In the top row, we see that as the distribution becomes steeper, the one minority sample is placed higher to receive proportional attention. In the bottom row, the flattest distribution requires a list of results were both classes are quite mixed. However, in the case of the steepest distribution, all elements of class B are placed in a block from rank two on to match the attention already given to class A by its representative at first rank.

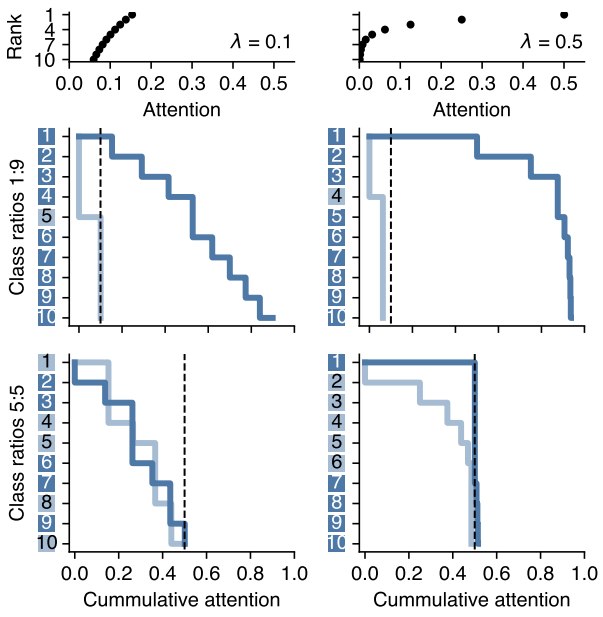


Figure 3: We generate rankings that satisfy our definition of attention fairness. Note that depending on the attention distribution function, fair rankings with the same proportions of classes differ.

5 CASE STUDIES

In this section, we apply the Viable-Λ Test to three different search engines: (1) gender fairness on a hiring site used by recruiters to search for candidates, (2) racial fairness on a dating site, and (3) political fairness on a search engine. As a disclaimer, we are only using these results to demonstrate a use of our metric on real-world data. We do not *not* claim that any of these are biased or unfair.

Ethics of Data Collection. While conducting our measurements we were considerate both of the services we collected the data from and of people who this data represents. All collected data is available to any person with an account on the corresponding service. We did not interact with any users of these services as part of the collection. Additionally, we minimized any impact on the operations of the services by using a low query intensity (at most one query every 30 seconds). Finally, we adhered to the usage quotas of Face++.[10]

Assumptions In our case studies, we assume that $W_R \sim Geom(\lambda)$. As discussed in Section § 4.2, $Geom(\lambda)$ meets the desired properties of an attention distribution, and thus serves as a reasonable starting approximate for the true W_R.

We used brute-force optimization over our one-dimensional parameter space D to check if $\exists \lambda \in D$ such that $d(E_R, \hat{p}) < \delta_{\max}$.

[7] The expected value of $Geom(\lambda)$ is $1/\lambda$. Since W_R is truncated, $E[W_R]$ asymptotically approaches this value for large n.
[8] We can use the confidence interval of μ to restrict the parameter space of other families of probability distributions by replacing $1/\lambda$ with the corresponding equation for $E(W_R)$.
[9] This approach is different from the eye-tracking studies mentioned in § 4.2. Instead of using eye-tracking heatmaps to construct the entirety of W_R, use a singular dimension of the data (namely $E[W_R]$) to inform a reasonable range of values for Λ.
[10] https://www.faceplusplus.com

5.1 Gender in Hiring

Our first cast study examines the ranking of job candidates on a resume search engine. Data was collected and made available to us by Chen et al. [9]. Using the recruiter's interface to the service, the authors queried 35 different search terms (such as "bartender", "electrical engineer", "laborer", "pharmacist", and "software engineer") in 20 US cities, resulting in 692 non-empty result lists. Of these results, we look at the 412 with length $n \geq 100$.

Modeling Gender Alignment L_R Chen et al. determined the gender of each candidate based on their given name (see [9] for details). The gender estimation is a real number describing the probability of a candidate being male between 0 (female) and 1 (male). The original authors dichotomized the data, assuming 'male' for probabilities >0.8, 'female' for probabilities <0.2, and omitted other profiles. These constituted 8% of all candidates. In this study, we use the gender alignment probabilities directly. We do neither omit ambiguous candidates, nor project alignments to their most-likely class.

Determining $\hat{p}$ In this study, the true population demographics for each query are unknown. As per Section § 4.3, we cannot choose an *implicit estimator* and instead leverage the vendor's data by choosing $\hat{p} \stackrel{\text{def}}{=} \bar{L_R}$ in each query.

Evaluating d and δ_{max}. $\hat{p}$ and E_R are both binomial distributions representing gender alignment. To check for statistically significant difference between them, we use the Z-test approximation for the binomial test. This $d(E_R, \hat{p}$ is our test statistic. We deem R_i unfair if $\nexists\lambda$ such that this test statistic is less than 1Z. This represents a 68% confidence that fair representation is impossible.

Viable-Λ in Hiring. In Figure 4, we run Viable-Λ for each list and plot the minimum attainable $d(E_R, \hat{p}$ for each R_i. About 92% of rankings can be considered fair at the 1Z threshold. Still, 6% of these lists appear biased against women and 3% appear unfair to men regardless of the distribution function. All rankings are deemed fair at the 2Z threshold

In Figure 5A, we demonstrate the effect of λ on $d(E_R, \hat{p})$. As per Section § 4.5, we assume within reason that $E[W_R] \in [.1n, .5n]$ in the context of this search engine. The horizontal axis depicts nine λ-parameter choices corresponding to a user viewing between 10% and 50% of all results on average. For all sampled values of λ, rankings under-represent women more often than men.

While the gender ratio is balanced in the dataset, the gender proportions vary widely between queries. We calculated the $\hat{p} = \bar{L}$ estimate for p for each query separately. Although the vast majority of rankings passed the Viable-Λ test, rankings tended to under-represent women more frequently for most values of λ.

5.2 Race in Dating

We used an author's personal account to query an online dating service's API. The script ran once every 15 minutes over the course of one week, collecting 672 lists each with length $n = 100$. We determined that each of these lists sampled from a pool 4,407 unique profiles. Even though we were running the same query repeatedly,

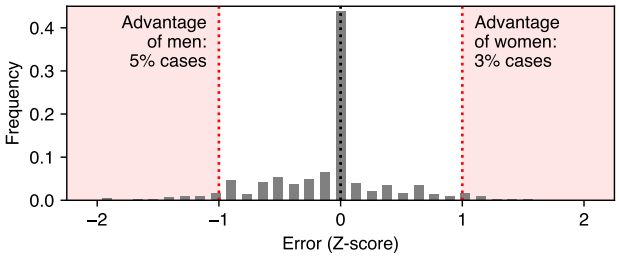


Figure 4: Best case scenario of error distribution among location-job pairs in the resume search engine. For each location-job query we attempt to find a Λ for which the ranking could be considered fair. We fail to do so in 8% of cases.

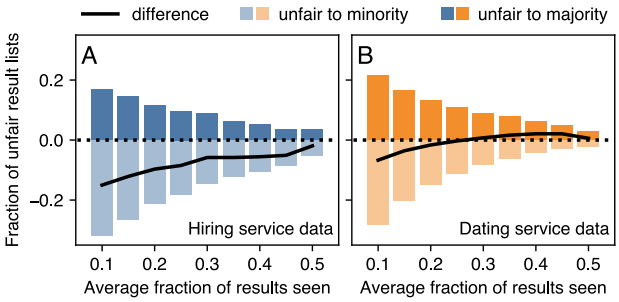


Figure 5: The more results a user sees, the higher the chance that the ranking can be considered fair (in extreme cases when all results are seen, the ranking is always fair). A) Regardless of λ, women are underrepresented more often than men in the job-location searches. B) Depending on the true value of λ, Black profiles may be underrepresented more frequently than non-Black profiles.

we observed significant *shuffling* and *churn* in the results, with some profiles being replaced more often than others.

Modelling Race Alignment. We used Face++ to infer the race of each person from their profile picture. Because the data is only used for illustration purposes, the precision of the race detection is not crucial.[11] In 813 (18.4%) profiles, Face++ detected > 1 face; among the identified profiles 2,411 (54.7%) were classified as white, 573 (13.0%) as Black, 534 (12.1%) as Asian, and 76 (1.7%) as Indian. For each profile photo, Face++ returns only the *most likely* race, not its certainty. One potential option is to model alignment as a categorical probability distribution among the five aforementioned classes. In doing so, we would use the χ^2 Independence Test to compute the statistical difference between E_R and $\hat{p}$. This would allow us measure the representation of all classes simultaneously. To simplify our illustrations, we project our alignment into a binomial distribution - i.e., Black vs. non-Black profiles.

[11] Face++ inferred a different gender than self-reported for 13% profiles identified as white, 11% profiles identified as Asian, and 8% of profiles identified as Black. Previous studies have shown high gender misclassification rates for Black women in particular [4]).

Determining $\hat{p}$, d, and δ_{max}. In this example, we ran the same query across each of our searches, yielding 4,407 unique results that meet the vendor's relevancy threshold. In our first approach, we chose $\hat{p} \stackrel{\text{def}}{=} \bar{L}$ on the set of all unique results. In doing so, we can build a 95% confidence interval for p; namely $\hat{p}$ is a *binomial sampling distribution* with $n = 4407$. Since we have a large sample size, our test statistic has a very small error threshold. We use the Z-test approximate for binomial test as our distance metric d again. Using this definition of $\hat{p}$, Viable-Λ showed that rankings unequivocally over-represented Black profiles. While Black profiles constituted only 13% of all unique profiles, the average ranking displayed 16% Black profiles.

Potential for Correcting Societal Bias. Black users have been shown to be disadvantaged in online dating [47]. Thus, it is possible that these profiles are over-sampled by this dating service to compensate for their lower click-through rates. Suppose that we want to measure how well rankings match the vendors' *manipulated* population demographics. Then in this case, p is the true percentage of Black profiles displayed by the vendor's ranking algorithm. At each rank k, we have a binomial sampling distribution of Black profiles with size $n = 672$ (i.e. the total number of searches). Thus, we can continue using the Z-test approximation for binomials as our distance metric.

Evaluating R_i Individually. We begin by evaluating each ranked list R_i individually. First, we find that for all values of the λ-parameter, the majority of lists can be considered fair, see Figure 5B. For the steepest distributions (smallest value of the λ-parameter) more result lists are unfair towards the Black users than towards non-Black users; when we assume an attention distribution function such that the users see on average 30% to 45% of results, there are more realizations in which Black users get more attention than proportional to their representation, and the situation equalizes for the least steep distributions. This effect is caused by the fact that white profiles appear on the first position in 54% of realizations, even though they only constitute 48% of the observed population on average, but for the rest of the high ranks, Black profiles are presented more often than the population estimator would indicate.

Evaluating R_i in Aggregate. Next, we evaluate the fairness of several realizations sampled in aggregate. A similar notion was proposed by Biega et al. [3]. As shown in Figure 6, even if each single ranking realization is unfair, the aggregate of multiple unfair rankings can be considered fair. Our metric can capture this because the alignment vector is not binary. We note that the more realizations included in the aggregate, the higher the fraction of fair aggregates, regardless of the assumed attention distribution function. Still, the steeper the function, the more runs are necessary to achieve a fair aggregate. In a hypothetical case of multiple rankings generated by a random (and, thus, unbiased [51, 54]) ranker, each rank will converge to contain a proportional representation of classes, and the ranking will be fair regardless of the assumed attention distribution function.

Summary. This case study highlights several important findings. The perception of the existence and even the direction of bias depends on the attention distribution. Furthermore, the bias can be corrected over time by reshuffling the results. Finally, our metric can accommodate population estimates that are different from the underlying populations for example to correct for societal biases.

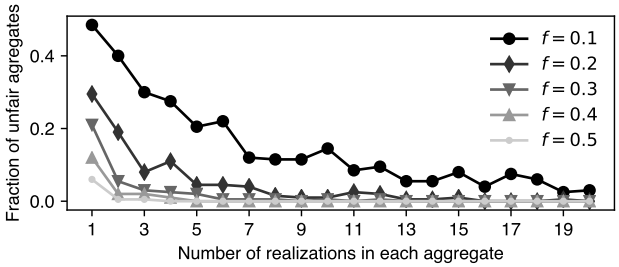


Figure 6: The investigated dating service returns a different set of results to the same query issued by the same person. As a consequence, while each ranking might be unfair, the results in aggregate are fair. The number of runs necessary to expect a fair ranking depends on the chosen distribution parameters.

5.3 Political Bias in Web Search

Our third case study differs from the previous two in that there is no protected group; instead, the *alignment* l_i is a proxy for political leaning of each item in the ranked list. The dataset we use was collected and made available to us by Robertson et al. [46]. There are two elements to the dataset: (1) search results and (2) partisan audience bias estimates. The first part comprises the first pages of results to 1,443 different web search queries. The second part maps 19,022 domains that appear in the search results to the bias scores on a liberal/conservative axis. Based on the tweets from registered voters, it assigns a number from -1 (only Democratic voters share content from that domain) through 0 (Democratic and Republican voters are equally likely to share content from this domain) to 1 (shared only by Republican voters). For example, `blacklivesmatter.com` scores -0.94, `en.wikipedia.com` scores -0.22, `dhs.gov` scores -0.01, `youtube.com` scores 0.13, and `catholics4trump.com` scores 0.98. Note that the score is assigned to a domain, not a specific webpage.

In this case study, we measure whether the aggregated partisan bias of search results is cancelled, given the attention distribution. Among the 1,443 searches in the dataset, we select three examples that best highlight the importance of considering attention distribution in the audit. In the interest of brevity, we omit the distance metric steps of Viable-Λ. We do, however, report the difference between the source biases weighted by the attention per rank and 0. Positive values indicate that a result list leans conservative; negative values indicate liberal lean.

Figure 7 presents the search results to three queries: "financial regulation", "obamacare continue", and "medicare reform", along with perceived partisan bias of each of these lists for different attention distribution functions. Note that in the first panel, small values of the λ-param correspond to a flatter distribution; as the λ-param grows, more attention is given to top results. The search

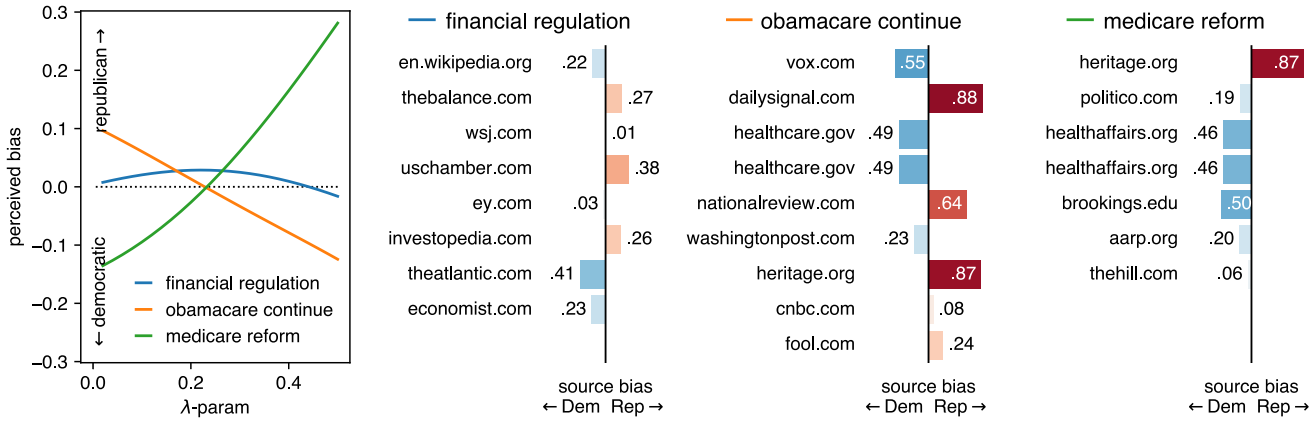


Figure 7: The perception of bias may depend on how the users of a service distribute their attention to the presented results. The neutral results for "financial regulation" form a list that appears neutral regardless of usage patterns. However, "obamacare continue" and "medicare reform" might appear partisan in either direction, depending on the attention distribution function.

engine might return multiple items from the same domain, therefore some domains appear multiple times in the list (for example `healthcare.gov` in Figure 7b).

The results in Figure 7 for "financial regulation" are neutral regardless of how attention is distributed since most of the results are apolitical. The results for "obamacare continue", on the other hand, lean republican: they feature three items from highly conservative sources. Still, because the first result is liberal-leaning—and among the first four results, three lean liberal—the result list will appear to be liberal-leaning overall if the attention distribution function is steeper. Finally, the results for "medicare reform" are almost exclusively from democratic-leaning sources. However, because the top result comes from a strongly conservative outlet, the list exhibits strong conservative-bias if the attention distribution function is steep. Thus, these examples illustrate how the shape of the attention distribution function can dramatically alter conclusions about the fairness/bias of ranked outputs.

6 LIMITATIONS

The Shape of W_R. While our framework allows for arbitrary families of W_R , we only considered the truncated geometric distribution in our examples. The results of the Viable-Λ test rely on an accurate model of human attention; thus further research into human perception and the quantification of the SEME would improve the basis of this metric.

Additionally, singlely-parametered W_R may not be sufficient in modelling expected attention. One important parameter is pagination; researchers have found that the CTR of the last result on a page getting more clicks than the pre-to-last [42], and each page introducing a disproportional drop-off of attention [26]. Furthermore, modern search engines often add variation into their search results. For example, Google search might display the actual content the user is seeking directly in the result page, or group results by type (i.e., "Sponsored", "Video", etc.).

Population Estimators. In this work, we derived the population estimators $\hat{p}$ directly from R using the $\bar{L}$ estimator. Hence, we assumed that the items the vendor chooses to show in the top k results are a proportional representation of all N potential results. It is likely, however, especially with large N, that they are not. For example, a real-world candidate ranking system employed by Amazon was shown to systematically rank women lower than men [11]. If we audited it and only had access to the top 100 results out of 1000, we would be likely to assume a $\hat{p}$ that underestimates the fraction of female candidates.

7 DISCUSSION

Studies have shown that swapping search results can cause significant changes in users' eye-gaze and click patterns. For example, after reshuffling Google Search results, unwitting users still tend to click on the top results, but some do shift more attention to lower ranks [35]. While we can use our metric to check if Λ exists such that representational parity is achieved (or—from the operator's point of view—verify that the results are fair given the known Λ), it is not guaranteed that altering results based on the metric's measurements will create fair rankings, since user behavior may change in uncertain ways. One possible way to create fair rankings is by means of a continuous, iterative process: the operator reshuffles the results to achieve parity under a measured Λ, users potentially change their behavior as a response, the operator updates the Λ estimate, and so on.

8 CONCLUSION

In this work we introduced a novel metric of group fairness in ranked lists, tying the measurement to the consumers' attention distribution. We showed how our approach could be used by auditors on three real world examples. Our results highlight the need for modelling attention specifically for the audited service: depending on the attention distribution function, the same list of results can appear biased both in favor and against the protected group. All code will be made publicly available upon publication.

ACKNOWLEDGMENTS

This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).
[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.
[3] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proc. of SIGIR*. ACM.
[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of FAT**. 77–91.
[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *Proc. of ICDM Workshops*. IEEE.
[6] Toon Calders, Asad Karim, Faisal Kamiran, Wesam Ali, and Xiangliang Zhang. 2013. Controlling Attribute Effect In Linear Regression. In *Proc. of ICDM*. IEEE.
[7] Toon Calders and Sicco Verwer. 2010. Three Naive Bayes Approaches For Discrimination-free Classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.
[8] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2017. Ranking with Fairness Constraints. *CoRR* abs/1704.06840 (2017). http://arxiv.org/abs/1704.06840
[9] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proc. of CHI*. ACM.
[10] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *Proc. of WSDM*. ACM.
[11] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[12] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For—How Search Informs Our Choice of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, M. Moore and D. Tambini (Eds.). 22.
[13] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. 2017. What Makes a Link Successful on Wikipedia?. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 917–926. https://doi.org/10.1145/3038912.3052613
[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proc. of ITCS*. ACM, 214–226.
[15] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.
[16] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1, 2 (November 2017).
[17] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proc. of KDD*. ACM.
[18] Google. 2016. Ads and analytics innovations for a mobile-first world. Google Blog. Accessed: 2018-10-29, https://www.blog.google/products/ads/ads-and-analytics-innovations-for-a-mobile-first-world/.
[19] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proc. of SIGIR*. ACM.
[20] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *ACM Conference on Fairness, Accountability, and Transparency*.
[21] Zhiwei Guan and Edward Cutrell. 2007. An eye tracking study of the effect of target rank on web search. In *Proc. of CHI*. ACM.
[22] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology For Direct And Indirect Discrimination Prevention In Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (2013), 1445–1459.
[23] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr.. In *Proc. of CSCW*. ACM.
[24] Eduardo Hargreaves, Claudio Agosti, Daniel Menasché, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. 2019. Fairness in online social network timelines: Measurements, models and mechanism design. *Performance Evaluation* 129 (2019), 15–39.
[25] Facebook Inc. 2018. https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Second-Quarter-2018-Results/default.aspx.
[26] Chitika Insights. 2013. The value of Google result positioning. *Online Advertising Network* (2013).
[27] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proc. of WSDM*. ACM.
[28] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Proc. of Conference on Computer, Control and Communication*.
[29] Faisal Kamiran and Toon Calders. 2010. Classification with No Discrimination by Preferential Sampling. In *Proc. of Machine Learning Conference of Belgium and The Netherlands*.
[30] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination Aware Decision Tree Learning. In *Proc. of ICDM*. IEEE.
[31] Faisal Kamiran, Asad Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *Proc. of ICDM*. IEEE.
[32] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Proc. of ECML PKDD*.
[33] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *Proc. of ICDM Workshops*. IEEE.
[34] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proc. of CHI*. ACM.
[35] Mark T Keane, Maeve O'Brien, and Barry Smyth. 2008. Are people biased in their use of search engines? *Commun. ACM* 51, 2 (2008), 49–52.
[36] Pauline T. Kim. 2017. Data-Driven Discrimination at Work. *William & Mary Law Review* 58 (2017).
[37] Tomasz J Kozubowski, Anna K Panorska, and Matthew L Forister. 2015. A discrete truncated Pareto distribution. *Statistical Methodology* 26 (2015), 135–150.
[38] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proc. of CSCW*. ACM.
[39] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *Proc. of KDD*. ACM.
[40] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81––97.
[41] Bennet B Murdock. 1962. The serial position effect of free recall. *Journal of Experimental Psychology* 64, 5 (1962), 482–488.
[42] Jamie Murphy, Charles Hofacker, and Richard Mizerski. 2006. Primacy and Recency Effects on Clicking Behavior. *Journal of Computer-Mediated Communication* 11, 2 (2006), 522–535. https://doi.org/10.1111/j.1083-6101.2006.00025.x
[43] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proc. of KDD*. ACM.
[44] Emily Rappleye. 2015. Gender ratio of nurses across 50 states. https://www.beckershospitalreview.com/human-capital-and-risk/gender-ratio-of-nurses-across-50-states.html.
[45] Matthew Richardson. 2007. Predicting clicks: Estimating the click-through rate for new ads. In *Proc. of WWW*. IW3C2.
[46] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction* 2 (2018).
[47] Christian Rudder. 2014. *Dataclysm: Love, Sex, Race, and Identity–What Our Online Lives Tell Us about Our Offline Selves*. Crown.
[48] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a preconference at the 64th Annual Meeting of the International Communication Association*.
[49] Piotr Sapiezynski, Valentin Kassarnig, Christo Wilson, Sune Lehmann, and Alan Mislove. 2017. Academic performance prediction in a gender-imbalanced environment. In *FATREC Workshop on Responsible Recommendation Proceedings*.
[50] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proc. of KDD*. ACM.
[51] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*.
[52] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment: Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proc. of WWW*. IW3C2.
[53] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Fairness Constraints: A Mechanism For Fair Classification. In *Proc. of FATML*.
[54] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proc. of CSCW*. ACM.
[55] Indre Zliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling Conditional Discrimination. In *Proc. of ICDM*. IEEE.

Unsupervised Topic Extraction from Privacy Policies

David Sarne
Bar-Ilan University
Ramat Gan, Israel
david.sarne@biu.ac.il

Jonathan Schler
Holon Institute of Technology
Holon, Israel
schler@hit.ac.il

Alon Singer
H-F & Co. Law Offices
Tel-Aviv, Israel
alon@h-f.co

Ayelet Sela
Bar-Ilan University
Ramat Gan, Israel
ayelet.sela@biu.ac.il

Ittai Bar Siman Tov
Bar-Ilan University
Ramat Gan, Israel
Ittai.Bar-Siman-Tov@biu.ac.il

ABSTRACT

This paper suggests the use of automatic topic modeling for large-scale corpora of privacy policies using unsupervised learning techniques. The advantages of using unsupervised learning for this task are numerous. The primary advantages include the ability to analyze any new corpus with a fraction of the effort required by supervised learning, the ability to study changes in topics of interest along time, and the ability to identify finer-grained topics of interest in these privacy policies. Based on general principles of document analysis we synthesize a cohesive framework for privacy policy topic modeling and apply it over a corpus of 4,982 privacy policies of mobile applications crawled from the Google Play Store. The results demonstrate that even with this relatively moderate-size corpus quite comprehensive insights can be attained regarding the focus and scope of current privacy policy documents. The topics extracted, their structure and the applicability of the unsupervised approach for that matter are validated through an extensive comparison to similar findings reported in prior work that uses supervised learning (which heavily depends on manual annotation of experts). The comparison suggests a substantial overlap between the topics found and those reported in prior work, and also unveils some new topics of interest.

CCS CONCEPTS

• **Information systems → Document topic models**; • **Theory of computation → Unsupervised learning and clustering**.

KEYWORDS

Topic modeling, unsuprevised learning, privacy policies

ACM Reference Format:
David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. 2019. Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317585

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317585

1 INTRODUCTION

A Privacy policy is a legal document that details the ways a website or a mobile application processes, collects, stores and shares end users' information, including their personally identifiable information. Typically, it also notes the purposes of these practices and the rights of the end users in that regard. Research on privacy policies has drawn much attention over recent years, for two main reasons. First, despite the importance of these documents as binding legal agreements, they are often not carefully considered by users, primarily due to their length, level of detail and the legal language used, all making them difficult to understand [1, 18]. Notably, the application of simplification techniques to reduce ambiguity and complexity of privacy policies has been found to have little or no effect on consumers' understanding of privacy policies, their willingness to share personal data, and their expectations about their rights [2, 26]. Second, the recent EU General Data Protection Regulation (GDPR) fostered substantial revisions to privacy policies, leading to various debates about the role of regulation in this important area [7, 13]. As such, much research has been carried out with the goal of understanding the structure, content and evolution of these policies.

In line with the growing trend of using AI technologies in LegalTech, there have been many recent reported efforts to use machine learning and text mining methodologies for analyzing privacy policies, aiming primarily at extracting and mapping the specific topics they address [10, 14, 16, 27]. Still, to date, all attempts to use machine learning and other AI-based tools for analyzing privacy policies utilized supervised-learning based techniques, which extensively rely on tagged repositories hence requiring substantial human effort for annotating the data [6, 23, 28].

In this paper, we suggest the analysis of corpora of privacy policies using topic modeling–an unsupervised learning technique. To that end, we use a cohesive framework which includes the construction of a targeted corpus, division into segments of interest, clustering into topics, and grouping. The use of unsupervised learning encapsulates many advantages. First, it enables analyzing any new corpus with a fraction of the effort required by supervised learning, which is highly useful for understanding specific classes of policies (e.g., based on genre, country of origin, culture). Second, it enables comparing privacy policies over time (e.g., before and after GDPR came into effect) to reflect changes in the topics addressed, compliance with the new regulations and evolution trends in general.

We apply the topic modeling framework over a corpus of 4,982 privacy policies crawled from the Google Play Store. We extract the most prevalent 36 topics encompassed in the corpus, enabling a thorough analysis of the structure, scope and focus of current privacy policies. In particular, the distribution of paragraphs over the different topics provides insight into what privacy policy drafters believe matters in terms of user privacy.

Naturally, the use of unsupervised learning calls for strict validation, as it is possible that the list of extracted topics is merely a subset of the actual list, focusing on limited areas of interest that are particularly structured. Therefore, as a means of validation the list of topics extracted is compared to a taxonomy found in Wilson et al [27], listing alternative set of privacy-policy topics, which is of wide use nowadays (e.g., see Harkous et al [10]). The comparison is carried out through manual mapping of the 36 extracted topics to those found in the referenced list. The results reveal a substantial overlap between topics, suggesting that the unsupervised topic modeling according to the proposed framework is highly effective and leads to a division into topics quite similar to that obtained with supervised learning techniques. Furthermore, several topics in our list do not have equivalences in prior work, possibly indicating new topics of interest that have been missed by supervised learning. In addition, several topics map to the topic identified in prior work possibly unfolding different aspects of the topic or suggesting finer-grained sub-topics of a rather general matter.

2 RELATED WORK

In 2017, the Global Privacy Enforcement Network (GPEN) published a report that evaluates the extent to which users control their personal information [8]. Based on a manual assessment of privacy notices, communications and practices of 455 websites and apps from various sectors by 24 data protection regulators from around the world, the report concluded that "there is significant room for improvement in terms of specific details contained in privacy communications". In this context, the development of automated tools for assessing privacy policies is highly desirable, and indeed in very recent years a growing number of studies suggested using machine-learning-based solutions [10, 27, 28], as well as other AI-related methods [15] to address this challenge.

Liu et al [15] suggested the use of Hidden Markov models (HMM) for privacy policy segmentation and for identifying whether a given pair of privacy policy paragraphs discuss the same topic. They found that HMM based segmentation methods performed on par with the lower half of human evaluators.Most others implemented supervised text-mining techniques for analyzing privacy policies. The use of supervised techniques requires a highly qualified annotated dataset for model construction. Such a dataset became available with the publishing of the OPP-115 manually annotated privacy policies corpus in 2016 by Wilson et al [27]. The corpus includes 115 privacy policies, annotated by domain experts using a designated annotation tool that was developed for this purpose. The tool offers a taxonomy based on 22 topics for the annotation purpose. The OPP-115 nurtured several studies that used supervised machine learning techniques for privacy policy analysis, using the data supplied as part of the corpus as a reference for validation. For example, Zimmeck et al [28] analyze thousands of mobile applications privacy policies by applying classifiers based on the OPP-115 corpus, trying to point, inter alia, inconsistencies between a mobile application and its privacy policy. Sathyendra et al [23] make use of OPP-115 corpus for training and evaluation of models used for automatic detection of opt-out provisions in privacy policies. Harkous et al [10] use OPP-115 for introducing an automatic framework for privacy policies (Polisis). The framework is based on a hierarchy of neural network classifiers, trained using the corpus data.

Common to all the above studies is their reliance on supervised learning. As discussed earlier, it requires either tagged repositories or investment of substantial human effort to annotate the data. Furthermore, this effort needs to be reinvested whenever analyzing specific classes of policies or attempting to analyze privacy policies over time in order to identify changes in a topic of interest. To the best of our knowledge, the only attempt to use unsupervised learning based techniques for analyzing privacy policies is the work of Ramanath et al [22]. Still, that work does not aim at topic modeling per se. Instead its focus is on identifying the points of transition between topics in privacy policies. Furthermore, that work evaluates the success of the proposed method using the subjective opinion of a human evaluator rather than by comparison to the findings reported in prior work.

Finally, we note that the GPEN report mentioned above also offers a taxonomy of topics related to privacy policies, generated by experts. These, however, focus primarily on users' control over personal information and include only 7 fairly wide topics (also called indicators in the report). All topics suggested there (except for a topic titled "automatic decision") are included in the OPP-115 list. Hence we base our comparison on the latter.

3 UNSUPERVISED LEARNING & TOPIC MODELING

Topic models are a type of statistical methods at frequent use for the discovery of abstract topics in a corpus [5, 9, 20]. Their underlying assumption is that there are groups of words (with different distributions) in the corpus, such that each represents a topic. A document in the corpus, contains those topics in different distributions. Topic modeling, which is an unsupervised machine learning method, attempts to extract the most probable distribution of words into topics through an iterative generative process which terminates upon convergence. The process does not require prior labeling or annotation of the documents. This makes it useful in complex or large data sets, where it is too cumbersome or complicated to extract topics manually [4].

Early topic models were described by Papadimitriou et al [21], followed by Hofmann's [11] Probabilistic Latent Semtnatic Analysis (PLSA) and Blei et al's [3] Latent Dirichlet Allocation (LDA) methods. The latter, which we use in this paper, is commonly used both for topic modeling in various domains (e.g. politics [9], psychology [20] and law [5]), and as a foundation for new topic model algorithms [12]. LDA assumes a sparse Dirichlet prior distribution over document-topic and topic-word distributions. During its iterative training phase, the algorithm learns and refines these distributions parameters. The resulting distributions are reflective of the assumption that topics usually use a relative small group of words, and a document contains a relatively small number of topics.

4 FRAMEWORK & IMPLEMENTATION

While there are various available "off-the-shelf" tools for topic modeling based on a corpus (e.g., mallet [19], graphlab (turi.com), familia (github.com/baidu/Familia)), the analysis of privacy policies is not trivial and requires several complementary capabilities, primarily related to content acquisition, content cleaning, segmentation and post-processing. In this section we propose a general framework for topic extraction of privacy policies. In addition, we provide the details and choices made in a specific implementation carried out in order to evaluate the effectiveness of unsupervised topic modeling of privacy policies, based on this framework.

Figure 1 provides an overview of our proposed design for topic modeling of privacy policies. While many of the components of our framework are commonly found in work dealing with corpus analysis, the synthesis we suggest is tailored for the specific case of privacy policies.

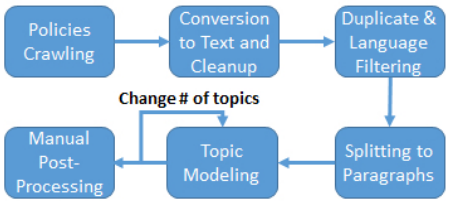


Figure 1: Overview of the topic modeling process.

Policies Crawling. To the best of our knowledge the only publicly available corpuses of privacy policies are the OPP-115 [27] and the ACL/COLING 2014 Dataset [15]. The rather small number of documents they contain (115 policies in OPP-115 and 1010 in the COLING Dataset) and the fact that the additional information they offer (i.e., labeling) is irrelevant for topic modeling using unsupervised learning makes them inappropriate for our purposes. Furthermore, any existing corpus is relevant to the time the policies it contains were collected. It may become obsolete as privacy policies are continuously being revised over time, especially in response to major regulatory changes, as in the case of GDPR [24]. Having the ability to construct new large and timely corpora is thus crucial. Here, crawling can be highly useful, as it offers full control over the number, properties and richness of the policies collected.

Taking the example of privacy policies of mobile apps, a new corpus can be constructed relatively easily, as a link to a relevant privacy policy document is provided as part of the information presented for any app in major app stores (e.g., Google Play, Apple's app store). Since these stores provide advanced search functionality, a corpus of a rather targeted set of policies can be constructed by automatically visiting the corresponding links and downloading the documents (that are stored on the app-developers' servers). Complementary meta-data related to the app itself (e.g., version, date, genre) can be downloaded as part of the process as it is stored in a rather structured manner in the HTML file.

For evaluation purposes we generated a corpus by crawling policy links from the Google Play Store and downloading the policies directly from the developers' servers. For diversity we selected policies that belong to 57 different Google Play categories. Overall, our crawler visited 24,071 app meta data pages during June 2018. Surprisingly, 19% of them did not have a privacy policy URL, leaving us with 19,458 URLs.

Conversion to Text, Cleanup and Filtering. Upon downloading a privacy policy, it should go through some pre-processing before added to the corpus. In many cases the first step is to convert the document into text (as it is formatted as an HTML file), removing tags, scripts and comments. This can be done using various off-the-shelf tools and libraries (e.g., Goose)[1]. Additional pre-processing include filtering of duplicate policies (as apps developed by the same company commonly use the same privacy policy), removal of policies in languages other than the intended one (e.g. by using NLTK or langdetect)[2] and removal of stop words (e.g. by using one of the stop words vocabularies commonly found in text mining tools such as GraphLab).

In our implementation, we removed 5,855 URL duplicates out of the 19,458 URLs crawled, leaving us with 13,603 URLs. We used the 'Goose' Python library to convert the documents to plain text and remove HTML meta-data. Further removal of text duplicates (identical privacy policies that were crawled from different URLs) and filtering out non-English language policies (using 'langdetect' Python library), led to a corpus of 4,982 documents, which served as the basis for our analysis.

Splitting to segments. Topic modeling requires the pre-processed policy to be divided into segments, each potentially dealing with a single topic. Here, one can use automatic tools for identifying points of transition between topics in privacy policies, such as the one suggested by Ramanath et al [22]. Alternatively, a common practice in topic modeling, which is particularly applicable for structured legal texts such as privacy policies, is to rely on the document's division into paragraphs as a means for segmentation [17]. The idea is to get to the smallest segments possible while keeping the essence of the text, bearing in mind that even if a topic is extracted based on paragraphs that are only part of a larger topic then later on it can be joined with topics representing other parts of the larger topic.[3] Furthermore, paragraphs can be easily identified and are generally seen as a self-contained unit of a discourse. Tools for dividing a document into paragraphs are abundant (e.g., NLTK's TextTiling Algorithm).[4]

In our implementation we preferred the division to paragraphs due to the above-mentioned advantages and the fact that the effectiveness of the alternative approach is somehow uncertain due to its reliance on subjective evaluation and the lack of comparison to an established set of topics. The process (carried out using TextTiling) resulted in 45,622 paragraphs. The average paragraph size was found to be 304 words, with standard deviation of 274 words and median of 248 words. From the latter set, we removed stop words using GraphLab's text analytics dictionary.

Topic Modeling and Assignment. Topic modeling requires to predefine the number of topics to be extracted. Trial and error may be required in order to set an effective number of topics, that will yield a set of cohesive and meaningful topics. The completeness and effectiveness of the set extracted can be measured in terms of the stylized statistics related to the division of paragraphs into

[1] github.com/GravityLabs/goose
[2] nltk.org/, github.com/Mimino666/langdetect.
[3] In fact, as demonstrated onwards, having several topics which are part of a larger topic enables finer-grained analysis.
[4] nltk.org/_modules/nltk/tokenize/texttiling.html

topics (e.g., mean, minimum, maximum, standard deviation), the distribution of confidence scores assigned to paragraphs within each topic (provided by the topic modeling tool), and possibly a comparison to labeled/existing data sets to validate the results. We note that while a large set of topics is likely to catch subtler aspects of privacy policies, it requires substantially greater amount of manual work for verification, merging and the construction of a topic hierarchy in order to identify higher level areas of interest in later stages.

In our implementation the topic modeling was carried out using GraphLab, a machine learning platform that supports LDA-based topic modeling. We set the number of topics to 100, as on the one hand it is large enough to represent the diversity of topics in a balanced way and on the other hand small enough to ensure that the manual merging of closely related topics is not overly tedious.

After completing the training phase, which took 600 iterations to converge, the LDA created clusters of words in each topic, as well as word probabilities in relation to each topic. Those probabilities were used to assign the most likely topic to each paragraph, by summing the word weights for all the paragraph words per topic, and choosing the topic that obtained the highest score.

Final Manual Processing. Topics obtained through topic modeling are identified by a collection of words. Typically, it is impossible to identify the substantive topic of each word collection by merely reviewing the words. A domain expert can thus be highly valuable, as she can sample some of the paragraphs assigned to each topic and identify its essence. She can also eliminate a topic if it is not cohesive. Topics that survive the elimination process can then be used for (manually) constructing a topics hierarchy, possibly pointing to higher level topics (areas of interest).

In our implementation we recruited as a domain expert a lawyer with 4-years of experience in privacy policies and commercial law. The expert went over each topic, reviewing a sample of at least 30 paragraphs, where paragraphs sampling from the ranked list (according to the confidence score assigned to each paragraph within the topic) followed the Padovan series [25], which provides a decent tradeoff between relevance (score) and sparsity. Based on the paragraphs sampled for each topic the expert provided a one sentence summary of the topic. 18 topics did not survive this process and were declared non-cohesive, leaving us with 82 topics.

Since the LDA model is designed in a manner that the distribution of the privacy policy paragraphs between the LDA model topics is uniform, some of the LDA models' topics cover legal issues that are identical. For example, in our case seven of the topics referred to personal information tracking technologies, using different forms and phrases or simply emphasizing different aspects of the same topic. Ideally, these paragraphs should be joined as they all apply to the same topic (legal issue) and they are not substantively distinguished. Still, the paragraphs are relatively diverse in terms of the words and structure used, such that even reducing the number of requested topics in the topic modeling process would not have grouped them all in a single topic. Therefore, whenever applicable, an expert should merge topics through manual processing. In our specific implementation the expert's merging process resulted in a total of 36 topics, each distinctly identified as dealing with one major legal issue.

5 ANALYSIS AND TOPICS VALIDATION

Table 1 lists the 36 topics that result from the expert's manual merger (third column). The table also specifies the number of topics that were merged into each topic (out of the original 82 topics) and the number of paragraphs that mapped to each topic (fourth and fifth columns, respectively). From the table we observe a substantial variation in the scope and level of detail of topics. Some topics are the result of merging several different original topics (e.g., the specification of the tracking technologies used for collecting personal information and the definition of what is personal information for that sake, each comprises seven original sub-topics) and have thousands of paragraphs mapped to them, possibly indicating the richness of the different ways in which they are formulated as well as the high level of detail they contain. However, most topics comprise a smaller number of original sub-topics and thus seem to show less variation in their formulation and possibly a narrower scope. Notably, three topics are exceptional in that they comprise only one or two of the original sub-topics and yet the number of paragraphs mapped to them is substantial: privacy policy update notification, contact information regarding personal information matters and user's option to opt-out of the privacy policy. This finding suggest that these are highly structured and standard topics that are commonly used.

Table 1 also maps each of the 36 topics in our list to the topics suggested by Wilson et al, which as discussed earlier were commonly used in prior work studies, thus enabling evaluating the validity of the topics mined with our framework.[5] The taxonomy used in Wilson et al's annotation system suggests (see also Figure 3 in Harkous et al [10]) a hierarchic presentation of topics where each topic in the top level of the hierarchy defines a high-level privacy category (first column in Table 1). The lower level (second column in Table 1) defines a set of privacy topics, with 22 such topics overall.

Notably, there is a substantial overlap in topics between the OPP-115 topics and those modeled in our implementation: topics in our list mapped into 17 out of the 22 topics listed in Wilson et al, which can be considered an important validation of our framework. Several of Wilson et al's topics are mapped by more than one topic from our list (four topics were mapped to two topics each and one to three topics from our list), possibly indicating the extraction of finer-grained topics through topic modeling. For example, within the topic of information type collected by 1st party, our topics distinguish between general information collected and financial personal information collected. As another example, the OPP-115 topic related to audience groups maps in our list into two topics that distinguish between the general population and the specific population of kids with respect to handling of personal information, a distinction that can be particularly important.

Five topics from the Wilson et al list are not present in our list (marked with grey background in the second column). These topics are related to information types collected by 3rd party, scope of access (for reviewing, editing, deleting, etc.), information type of data retained, user choice related to policy change and practices that are not covered in the policy. We note that each of those belongs to a different area and all other sibling topics of that level are adequately

[5] The mapping was carried out manually by the domain expert.

Table 1: Mapping of topics to those suggested by Wilson et al [27].

OPP-115 Category	OPP-115 Topic	Topics in Our Unsupervised-Learning Implementation	Merged	Paragraphs
1st party collection	Collection mode	Personal information tracking technologies	7	3674
		Personal information collection methods	2	890
	Information type	Collected personal information description	3	1237
		Collected financial personal information description	1	655
	Purpose	Personal information collection purposes	4	1706
3rd party collection	Action	Personal information shared with 3rd Parties	6	2915
		Personal information collection directly by 3rd parties through app	2	584
	Information type	-	-	-
	Purpose	Forced disclosure of personal information (e.g. court order)	3	1366
	-	Disclosure of personal information by user directly to 3rd parties	1	571
Access. Edit, Delete	Access scope	-	-	-
	Access rights	Users right to view/update/delete their personal information	5	2610
Data Retention	Retention period	Personal information retention period	1	456
	Retention purpose	Purpose of personal information collection	1	429
	Information type	-	-	-
	-	Collected technical related personal information description	2	1221
		Collected geographical data	1	593
		App does not collect personal information	1	507
		Implication of granting permission to app to collect personal information	1	468
Data Security	Security measure	How personal information is protected	1	291
Int. and Specific Audience	Audience group	Personal information regulations	4	1801
		Geographical transfer of personal information	3	1609
		Personal information of kids	1	885
Do Not Track	Do not track policy	Do Not Track	1	321
Policy Change	Change type	Privacy policy updates	1	322
	User choice	-	-	-
	Notification type	Privacy policy update notifications	1	1071
Other	Introductory	What is personal information	7	3408
	Contact information	Contact information regarding personal information matters	1	1013
	Practice note covered	-	-	-
	-	Paragraphs that are irrelevant to privacy	3	1143
		General legal provisions (e.g. Limitation of liability/governing Law/jurisdiction/disclaimers)	5	2457
		Collected non-personal information	4	1150
		Details about creation of an account and related password	1	440
		App developer representations regarding compliance with personal information regulations	1	335
Choice Control	Choice type	Clarification regarding user's own decision to share their personal information	1	120
		User option to opt-out	2	1585
	Choice scope	Notice that user's acceptance of personal information processing can be withdrawn	1	534
	-	Sharing of personal information by user and related risks	1	667
		Notice about user's explicit grant of right to use their personal information by the app	1	597
		Warnings from unwanted disclosure of personal information	1	443

mapped. Thus, the absence of the five topics from our list does not result in missing a complete area of interest, but rather in lack of specification of a certain aspect that possibly exhibits high variation in the way it is formulated in privacy policies. Interestingly, there is no correlation between the number of paragraphs mapped into a topic (hence possibly its importance or complexity) in our list and its appearance or absence in the other list.

On the other hand, 13 topics from our list do not map to any of the topics in Wilson et al (marked with grey background in the third column). Some of these topics do not directly address privacy issues (e.g., Collected non-personal information, General legal provisions (e.g. Limitation of liability/governing Law/jurisdiction)), yet others certainly unfold new considerations and aspects that have been missed in prior work, e.g., risks related to sharing personal information by the user, privacy policy updates, and disclosure of personal information by the user directly to 3rd parties. Eight of the specific topics we found that do not map to areas of interest in Wilson et al's taxonomy, suggest new topics of interest within the specific area, and five do not map to an existing area, although they do not appear to warrant a new area of interest of their own.

Finally, we note that our topic-modeling-based analysis yields insights related to the importance and weight of the different levels of hierarchies (areas) suggested in prior work as well as to the extent of details they exhibit—the greater the number of topics associated with each category and the number of paragraphs mapped to them, the greater the breadth of aspects the category comprises and the level of detail or complexity it entails. These notions are illustrated in Table 2 through a comparison to the set of expert-derived areas of interest provided in Wilson et al. The table depicts the percentage of paragraphs mapped into each area of interest out of the total number of paragraphs in our corpus (second column), based on the mapping given in Table 1. The third column is the percentage of annotations found in the OPP-115 corpus mapped to each area out of the total number of annotations there (see Table 2 in Wilson et al).

Table 2: OPP-115 [27] topics and the number of topics from our analysis corresponding to each.

OPP-115 Category	% of Paragraphs	% annotations in OPP-115
Other	23.0%	15.4%
1st Party Collection	18.7%	38.5%
3rd Party Collection	13.6%	22.5%
International and Specific Audience	12.4%	4%
Data Retention	10.9%	1.6%
Choice Control	9.9%	7.7%
Access, Edit, Delete	6.5%	3.2%
Policy Change	3.5%	2.4%
Do Not Track	0.8%	0.4%
Data Security	0.7%	4.3%

We learn from the table that even though the distribution of segments into topic areas is somewhat different, in general areas that received greater weight in the manual annotations of OPP-115 were identified as the major areas of interest also by our unsupervised topic modeling. With respect to the most minor areas of interest, the two methods provide almost the same weight assessment. A

substantial difference with the major-weight areas derives from the difference in the "Other" topic, arguably due to the new topics we found in our topic-modeling based analysis, which were naturally assigned into this category. Hence we are able to provide a good estimation of the weight and breadth of different privacy-policies-related categories without the extensive human labor that is required for labeling.

6 DISCUSSION AND CONCLUSIONS

The encouraging results related to our framework validation reported in the previous section suggests that topic modeling based on unsupervised learning can be highly effective in extracting topics of interest of privacy policy corpuses. The implications for regulators are many, as discussed throughout the paper. This method provides a means for monitoring the evolvement of privacy policies over time and for reflecting changes in their content, in terms of the topics being addressed, given a specific set of regulations or following a regulatory change.

To the best of our knowledge, our work is the first to propose, implement and evaluate an unsupervised topic modeling framework for analyzing privacy policies topics of interest. In addition to validating the findings the comparison to previous work provides a somewhat finer-grained list of topics under each high-level area of interest. Additionally, we find a set of topics that were not proposed through previous applications of supervised analysis. One possible explanation for the newly added topics is the recent coming into effect of GDPR. It points to the importance of the proposed unsupervised method and highlights its usefulness and contribution in an ever changing regulatory environment.

While unsupervised learning does not require the human labor necessary for annotating the data in supervised learning, it does require some manual post-processing by a human expert to verify the coherence of the topics mined, summarizing the content of the different topics and merging them whenever applicable. We emphasize that this human effort is several orders of magnitude smaller than the effort required for annotating data for supervised methods.[6]

We hope that the success of the specific implementation reported in the paper and the public availability of the corpus (to which we keep adding policies and newer versions of existing policies for purposes of comparative analysis) will facilitate further insights and contribution. Other directions for future work include the analysis of differences in privacy policies across domains and over time. In addition, we plan to evaluate advanced methods for document segmentation, beyond current paragraph-based approach, that will better reflect changes in the topic discussed.

ACKNOWLEDGMENTS

This research was partially supported by the ISRAEL SCIENCE FOUNDATION grant No. 1162/17.

REFERENCES

[1] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies* 43, 1 (2014), 1–35.

[2] Omri Ben-Shahar and Adam Chilton. 2016. Simplification of privacy disclosures: an experimental test. *The Journal of Legal Studies* 45, S2 (2016), S41–S67.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3 (2003), 993–1022.

[4] Zhiyuan Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *International Conference on Machine Learning*. 703–711.

[5] Hyo Shin Choi, Won Sang Lee, and So Young Sohn. 2017. Analyzing research trends in personal information privacy using topic modeling. *Computers & Security* 67 (2017), 244–253.

[6] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In *Proc. of WPES*. 91–96.

[7] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz. 2018. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. *ArXiv e-prints* (2018).

[8] GPEN. 2017. GPEN Sweep 2017 - User Controls over Personal information.

[9] Derek Greene and James P Cross. 2017. Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Political Analysis* 25, 1 (2017).

[10] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[11] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of UAI*. 289–296.

[12] Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*. 577–584.

[13] T. Linden, H. Harkous, and K. Fawaz. 2018. The Privacy Policy Landscape After the GDPR. *ArXiv e-prints* (Sept. 2018). arXiv:cs.CR/1809.08396

[14] Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016. Modeling language vagueness in privacy policies using deep neural networks. In *AAAI Fall Symposium on Privacy and Language Technologies*.

[15] Fei Liu, Rohan Ramanath, Norman M. Sadeh, and Noah A. Smith. 2014. A Step Towards Usable Privacy Policy: Automatic Alignment of Privacy Statements. In *COLING*. ACL, 884–894.

[16] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards Automatic Classification of Privacy Policy Text. CMU-ISR-17-118R, CMU-LTI-17-010 (June 2018).

[17] Yue Lu and Chengxiang Zhai. 2008. Opinion Integration Through Semi-supervised Topic Modeling. In *Proc. of WWW*. 121–130. https://doi.org/10.1145/1367497.1367514

[18] Florencia Marotta-Wurgler. 2012. Does Contract Disclosure Matter? *JITE* 168, 1 (2012), 94–119.

[19] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002). http://mallet.cs.umass.edu.

[20] Kate Niederhoffer, Jonathan Schler, Patrick Crutchley, Kate Loveys, and Glen Coppersmith. 2017. In your wildest dreams: the language and psychological features of dreams. In *Proc. of CLPsych*. 13–25.

[21] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent Semantic Indexing: A Probabilistic Analysis. In *Proc. of PODS*. 159–168. https://doi.org/10.1145/275487.275505

[22] Rohan Ramanath, Fei Liu, Norman M. Sadeh, and Noah A. Smith. 2014. Unsupervised Alignment of Privacy Policies using Hidden Markov Models. In *ACL (2)*. 605–610.

[23] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the Provision of Choices in Privacy Policy Text. In *Proc. of EMNLP*. 2764–2769.

[24] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2018. Analyzing Privacy Policies Using Contextual Integrity Annotations. *arXiv preprint arXiv:1809.02236* (2018).

[25] I. Stewart. 1996. Tales of a Neglected Number. *Scientific American* 274 (June 1996), 102–103. https://doi.org/10.1038/scientificamerican0696-102

[26] Lior Jacob Strahilevitz and Matthew B Kugler. 2016. Is Privacy Policy Language Irrelevant to Consumers? *The Journal of Legal Studies* 45, S2 (2016), S69–S95.

[27] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proc. of ACL*. 1330–1340. https://doi.org/10.18653/v1/P16-1126

[28] Sebastian Zimmeck, Lieyong Zou Ziqi Wang, Bin Liu Roger Iyengar, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated analysis of privacy requirements for mobile apps. In *Proc. of NDSS*.

[6] For example, the OPP-115 corpus includes 23,194 annotated data practices, 128,347 annotated attributes and 102,576 annotated text spans 102,576, all performed manually.

Fairness in Social Influence Maximization

Ana-Andreea Stoica
Columbia University
astoica@cs.columbia.edu

Augustin Chaintreau
Columbia University
augustin@cs.columbia.edu

ABSTRACT

Algorithms for social influence maximization have been extensively studied for the purpose of strategically choosing an initial set of individuals in a social network from which information gets propagated. With many applications in advertisement, news spread, vaccination, and online trend-setting, this problem is a central one in understanding how information flows in a network of individuals. As human networks may encode historical biases, algorithms performing on them might capture and reproduce such biases when automating outcomes.

In this work, we study the social influence maximization problem for the purpose of designing fair algorithms for diffusion, aiming to understand the effect of communities in the creation of disparate impact among network participants based on demographic attributes (gender, race etc). We propose a set of definitions and models for assessing the fairness-utility tradeoff in designing algorithms that maximize influence through a mathematical model of diffusion and an empirical analysis of a collected dataset from Instagram. Our work shows that being feature-aware can lead to more diverse outcomes in outreach and seed selection, as well as better efficiency, than being feature-blind.

CCS CONCEPTS

• **Theory of computation** → **Network flows**; • **General and reference** → *Empirical studies.*

KEYWORDS

social influence; graphs; social networks; seeds

ACM Reference Format:
Ana-Andreea Stoica and Augustin Chaintreau. 2019. Fairness in Social Influence Maximization. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317588

1 INTRODUCTION

Social influence maximization has been a widely studied problem in online networks, having impactful applications in advertisement campaigns, viral online content, news propagation, disease spread, and many others. In spreading an idea, product, or technology through a social network, the network structure plays a crucial role in the efficient propagation of such a process: who *has* the information and who they are *connected* to determines who that information reaches. Thus, the status of being 'early-adopter', who either adopts the desired product out of their own will or receives it for free, is a privileged position for their social status and their direct impact in the diffusion process. The early-adopters promote the content to their friends, who in turn may or may not adopt it and continue the process, resulting in a cascade.

In maximizing the outreach, choosing the early-adopters strategically is crucial. Many algorithms have been proposed in finding an optimal set, from a greedy choice to leveraging the network structure through degree or distance centrality. In theory, such algorithms are blind to demographics, as they typically assess the chance for a given individual to be an early-adopter solely based on their position in a network. In practice, however, properties of social networks also encode historical biases and gender artifacts that algorithms can reinforce.

In this paper, we aim to understand the mechanisms behind information diffusion in social networks and to design algorithms that maximize social influence in a *fair* way. Fairness in machine learning and in graph algorithms is becoming an increasingly popular field in the computer science, sociology, policy, and law, as recent work shows the effect of automated algorithms in mirroring or amplifying bias in human datasets. This is an even more urgent problem in the case of relational data, where connections between individuals may be used as proxies for location, income, social status, and can thus be used by algorithms to exacerbate inequality between different demographic groups.

To our knowledge, however, fairness in social influence maximization has yet to be defined, nevertheless solved. Thus, we begin by analyzing the state-of-the-art algorithms for information diffusion and the underlying graph models, following by adapting various fairness definitions used in classification settings. We perform an empirical analysis using a collected dataset from Instagram, analyzing the effect of different seeding heuristics on the community structure. Furthermore, we develop a mathematical model to formally compare strategic and non-strategic seeding and quantify their effect on disparate outcome among different populations of a network, using as a diffusion process the independent cascade model. We conclude by formulating a way of tackling fairness in the social influence maximization problem, arguing that a strategic heuristic that is feature-aware can be fairer and more efficient.

2 RELATED WORK

Previous literature tackles this problem as an optimization problem by finding a submodular function that models the reach and optimizing it to find the best seed set. The pioneering work of social influence algorithms by [11, 12] in 2003 kicked off by proving that submodular optimization extends to a large collection of social influence processes. Amongst other more robust heuristics and efficient implementations [6], more recent papers extended such methods to

This material is based upon work supported by the National Science Foundation under Grant No.1254035 and 1514437.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317588

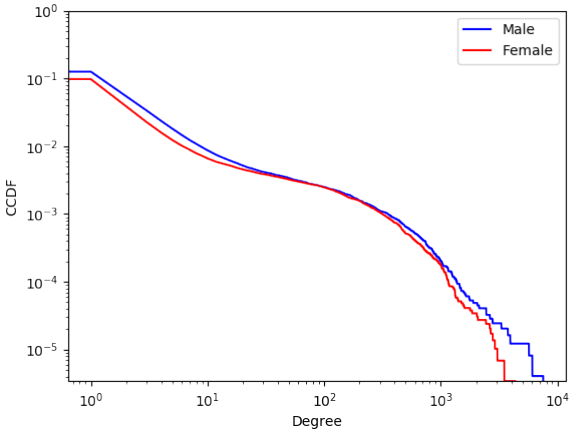


Figure 1: Complementary cumulative distribution function of the degrees in the Instagram graph, plotted for the males (blue) and the females (red).

deal with uncertain networks through online adaptive queries [18], and recently established the importance of exploiting community structure to best direct influence [4]. In this paper, we focus on the algorithmic aspect of the problem, analyzing current algorithms for maximizing influence and their implications on social bias.

A recent line of work focuses on fairness in algorithmic design for human data, with a large focus on classification tasks. From the classic example of predictive policing algorithms that reproduce bias in the data [1, 5, 22], the field of fairness in machine learning developed several ways of measuring and defining *fairness*, from defining statistical constraints on the output based on the particularities of the input data [8, 14], to pre-processing biased data [9, 10, 23], and to identifying the causal relationships between features in the data that lead to bias [13, 16]. Measurement techniques have also been developed in order to understand bias in massive data [21].

Amongst these efforts, two schools of thought emerge, one that argues for a feature-blind strategy in order to not reinforce biases from input data, and one that argues for a feature-aware strategy that mitigates bias after learning it from data. While both have applications in different contexts, an argument that [7] makes is that feature-blindness might lead to amplifying bias through confounding factors. In our work, we explore these two views and show indeed that a feature-blind strategy for social influence may cause unfair outcomes. In order to understand the effect of strategy in seed selection, we ask the question: what does it mean to be fair in information diffusion?

As mentioned in the introduction, we are interested in the effect that social influence maximization has on the different communities present in the network. As such, we may define the following:

(1) *Fairness for early-adopters:* Given a network with communities $C_1, C_2, \cdots C_k$, a social influence maximization algorithm is fair in choosing a seed set S if the proportion of all communities in the seed set is equal,

$$\frac{\mathbb{E}|\{u \in S | u \in C_i\}|}{|C_i|} = \frac{\mathbb{E}|\{u \in S | u \in C_j\}|}{|C_j|}. \quad (1)$$

(2) *Fairness in outreach:* Given a network with communities $C_1, C_2, \cdots C_k$, a social influence maximization algorithm is fair in the outreach it achieves if the cascade reaches all communities in a calibrated way,

$$\frac{\mathbb{E}(|\{u \text{ adopted } | u \in C_i\}|)}{|C_i|} = \frac{\mathbb{E}(|\{u \text{ adopted } | u \in C_j\}|)}{|C_j|} \forall i, j. \quad (2)$$

This notion can ensure that the product or the information being spread reaches the population in a calibrated manner, reaching the same percentage of each community. This is particularly important in news spread, where manipulating distribution of news can lead to misinformation and to amplifying an echo chamber effect, and maybe even more compelling for public health concerns..

Both definitions are in essence a form of calibration or statistical parity (as defined in [14]), and we argue that both are necessary: while the outreach definition ensures that the product or idea reaches all communities equally, early-adopters have a special role in the diffusion process. Having early access to products may establish their role as influencers in the network and allowing them to acquire social capital and trust, as well as to leverage that idea or product for financial gain. A classic example is being an early adopter in Bitcoin, where any user who had access to early mining had the chance of exponentially increasing their profits even when individual mining stopped being profitable. Similarly, influencers on Instagram get paid by companies to promote products. Due to the popularity of the users, the product is more likely to be seen or adopted, and thus being an early-adopter has its perks.

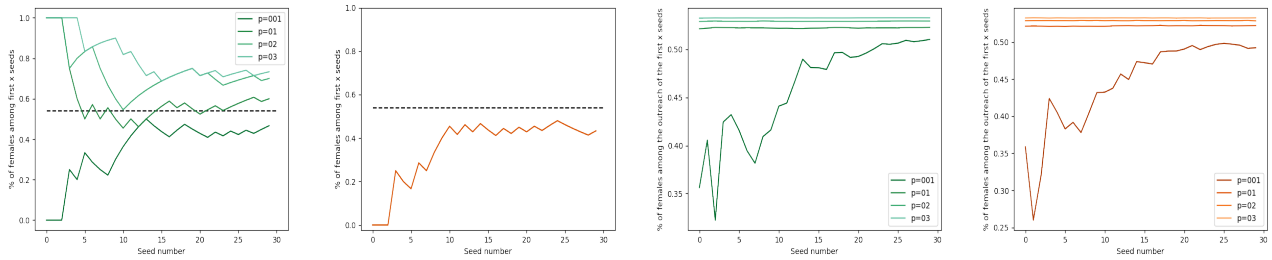
3 EMPIRICAL ANALYSIS

We performed experiments on a collected dataset from Instagram, consisting of public profiles, photos, usernames, likes and comments received on photos, and so on. As Instagram is one of the main hubs for advertisement purposes, with users continuously promoting paid content through the photos or descriptions they are posting, this network provides a comprehensive view of how information diffusion occurs.

3.1 Data Collection

In collecting the data, we used the Instagram API to crawl public profiles, starting from the founder of Instagram, Kevin Systrom. For each such public photo, we recorded the author, the number of likes and comments, and who liked and commented on the item. In total, we collected 115, 796, 284 photos over multiple months of 2014 and 2015, which amounted to 539, 023 different users for whom we could infer gender from their names, as described below. While the data might not encompass the full diversity of Instagram users, it is sufficient to obtain statistically significant results most of the time.

Gender inference. We inferred the gender of the users from their first name, adapting the method from [17] that leverages first names and social security data. Filtering for names with less than 50 occurrences for either gender, we obtained a set of 32, 676 unique first



(a) Greedy (b) Degree (c) Greedy (d) Degree

Figure 2: Percentage of females among chosen seeds for the greedy algorithm (a) and the degree algorithm (b), and among the outreach for the greedy algorithm (c) and the degree algorithm (d) for different values of p .

names that we used to label about half a million users. The gender proportions are 45.57% males and 54.43% females.

The graph of likes and comments. In creating a network structure out of these aggregated users and their information, we used the likes and comments they left on eachother's posts as interactions. Whenever someone gave a like or comment on a photo, we created an edge from the originator to the recipient, ending up with **640,211 edges** between **539,023 nodes**.

We will be using this graph for the purpose of our analysis. The dataset consists of many more users and interactions for which we could not determine gender.

A first analysis reveals a disparity in the degree distribution for the two genders. As Figure 1 shows, although males are in minority, they have higher degrees (number of interactions) than females, gap that increases for highly connected nodes.

3.2 Influence maximization algorithms

In studying strategic choices of seeds versus label-blind choices, we focus on two main algorithms, the greedy algorithm as a strategic one and the high-degree algorithm as one that ignores the labels of the nodes, and their effect on a bipopulated network represented by the Instagram dataset.

Greedy algorithm. The greedy algorithm is defined as choosing the seeds sequentially, each time choosing the node that gives the best marginal increase in the resulting cascade (outreach), until we reach a seed set of a predefined size. [11] shows that this provides a $1 - 1/e$-approximation to the best choice of seeds overall. In our implementation, we choose the seedset size to be equal to $k = 30$ nodes, and assume the independent cascade model for information diffusion. Since this model is probabilistic, at each step of the process we choose the node that gives the best marginal outreach averaged over 1000 realizations, choosing conducting edges pseudo-randomly while varying conducting probability p. This algorithm is strategic since the marginal contribution of a node might be dependent on the history of chosen seeds and their labels.

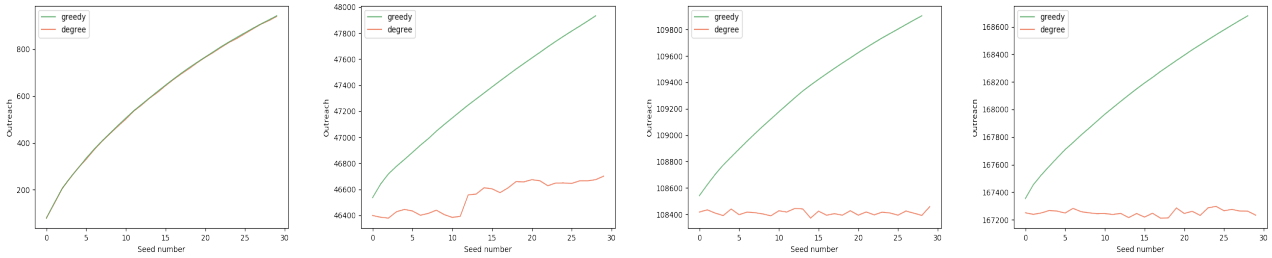
Degree algorithm. Widely used in sociology literature [20], this heuristic chooses nodes as seeds in decreasing order of their degree in the graph and is commonly known as a degree-centrality heuristic. The node selection is done independently of the labels and of the conducting probability p, yet measuring outreach is not and requires again an approximation technique due to the intractability of the problem. We again average over 1000 realizations just as above.

Fairness in seed selection. Figure 2 (a-b) shows the percentage of females in the choice of seed sets for each of the above algorithms. For the greedy algorithm, we notice that for $p = 0.01$ the fraction of females is lower than their fraction in the network, 54%, for small seedset sizes, and it slowly converges towards their true fraction as the seedset increases. However, for $p = 0.1, 0.2, 0.3$, the seedset starts by being *exclusively* female and slowly converges towards a more equilibrated gender ratio as more seeds are added.

For the degree algorithm, men are at the top with higher degrees than women, so the seed selection captures this in Figure 2 (b). We observe that even after choosing $k = 30$ seeds, women are under-represented, given their true fraction of 54%. Our results are consistent with varying the seedset size. Given that the degrees of women are not only lower at the very top but in general (Figure 1), this disparity will continue as the number of seeds increasing, deeming the degree algorithm generally unfair in representing one of the populations. It is generally difficult to assess whether this algorithm reinforces or simply perpetuates a present inequality, yet it raises the question of whether and how can we leverage the degree information in a way that is both fair and efficient.

Fairness in outreach. Figure 2 (c-d) shows the percentage of females in outreach as the seedset size increases up to $k = 30$, for the greedy and degree algorithms. Consistent with the measurements for seedset composition, the percentage of females in outreach converges to their true fraction of 54% as p increases from 0.1 to 0.3. For $p = 0.01$, however, both algorithms fail to reach a proportionate number of women in the network, the degree algorithm performing slightly worse than the greedy one.

Such results give a sense of the trade-off between fairness as diversity and efficiency. Figure 3 shows that the greedy algorithm achieves far better outreach than the degree heuristic for $p = 0.1, 0.2, 0.3$, and more or less the same for $p = 0.01$. Coincidentally, for $p = 0.01$, both algorithms prefer male over female although the male community is in minority. This is even the more meaningful as the first seed generally carries most of the marginal gain for influence for $p = 0.1, 0.2, 0.3$ (see Figure 3 (b-d)), and so



(a) (b) (c) (d)

Figure 3: Performance in terms of average outreach for the greedy algorithm (green) and the degree algorithm (orange) for $p = 0.01$ (a), $p = 0.1$ (b), $p = 0.2$ (c), $p = 0.3$ (d).

$p = 0.01$ represents a case where the influence of each subsequent seed really affects the outcome.

Thus, for such a case being strategic can also mean being more fair. As [4] argues in a similar fashion, in the case of an independent cascade, diversifying between communities gives a better outreach to nodes that otherwise would not be reached.

While the choice of the conducting probability p may differ in different scenarios and should be measured when such data is available, clearly the case in between $p = 0.01$ and $p = 0.1$ is the most notable one, since beyond that the first seed carries a disproportionate amount of the weight.

However, even in such cases being strategic doesn't necessarily mean implementing a greedy heuristic. Due to the computational difficulty of the greedy algorithm and the potential availability of other metrics in the data (such as degree, for example), it is important to understand the implications of such metrics. While an available metric might lead to a label-blind heuristic, that might also lead to disparate outcomes. As we have seen, a degree heuristic will always be unfair in the seed selection in a network whose populations follow different degree distributions (in this case, a majority population gets under-represented because of lower degrees).

In exploring this avenue of strategic seeding in a fair manner we turn to a theoretical model for understanding the cascading process.

4 THE LIMITATIONS OF LABEL-BLIND SEEDING

An established model of network growth that encapsulates clustering dynamics and a power-law degree distribution is the preferential attachment model. Prior work [2, 3, 15, 19] stresses the applicability of such a model into describing real-world network formation, having the caveat of a theoretical understanding of its dynamics. For the purpose of this study, we use the preferential attachment model with two communities and homophily, as defined below.

Preferential attachment model with homophily:

- *Minority-majority partition*: every timestep a new node appears and chooses a label, red (R) w.p. r and blue (B) w.p. $1 - r$, where $0 < r < 1$.
- Every timestep the new node connects to some other node according to *preferential attachment*, meaning that it choses a node in the network with probability proportional to that node's degree. In our model each new node forms $d > 1$ connections.
- *Homophily:* if the two nodes have the same label, connect, otherwise accept the connection w.p. ρ, where $0 < \rho < 1$.

This model results in a community structure (nodes of each label are homophilic), and when the two communities differ in sizes, so do their degree distributions. Indeed, [3] show that the degree distribution of this network follows power law for the R and B nodes with coefficients

$$\beta(B) = 1 + \frac{1}{\frac{r\rho}{2\alpha+2(1-\alpha)\rho} + \frac{1-r}{2\alpha\rho+2(1-\alpha)}}, \tag{3}$$

$$\beta(R) = 1 + \frac{1}{\frac{(1-r)\rho}{2\alpha\rho+2(1-\alpha)} + \frac{r}{2\alpha+2(1-\alpha)\rho}}, \tag{4}$$

where α denotes the fraction of edges that emerge from red nodes and the total number of edges.

When the proportions of R and B nodes are different ($r < 1/2$), we know that $\beta(R) > 3 > \beta(B) > 2$ and that a so-called glass ceiling effect is present (the difference in degrees for the two communities increases for more well-connected nodes). Our Instagram dataset follows a two-community structure as well (men and women), and while men are in minority, they have higher degrees than women, with a degree distribution resembling a power-law with different coefficients (Figure 1).

Independent cascade: On top of this model we simulate the independent cascade mechanism, where each node that receives the information/product/etc will adopt and promote to his friends with probability p, for $0 < p < 1$, in order to achieve an optimal seedset of size $k(n)$, a function that depends on the network size n.

Strategic seeding: In the social influence maximization problem the task is to find a set of "seeds" in the network starting from which information diffusion has maximal outreach. In the independent cascade model, the greedy algorithm is shown to approximate the optimal choice of sees [11], yet it is NP-complete and thus very hard to analyze. However, in many cases, choosing the nodes of top degree achieves a similar effect as the greedy algorithm. For the purpose of this study we assume that baseline strategic seeding

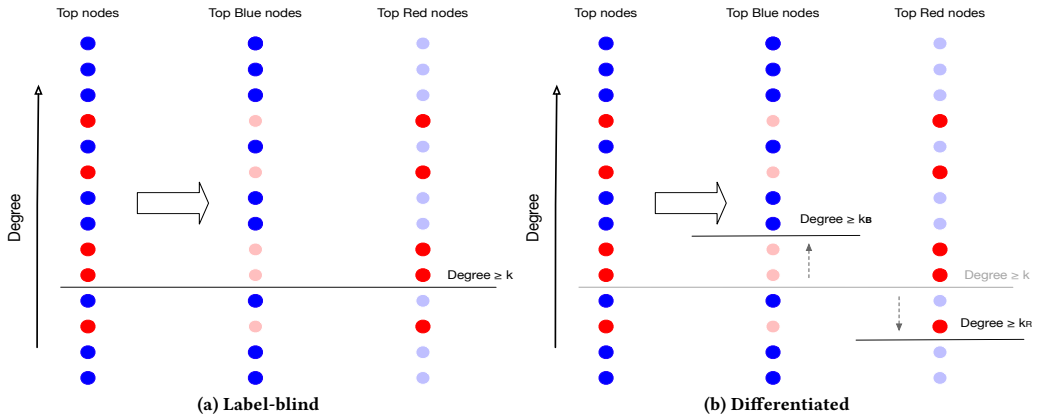


Figure 4: Model of choosing seed nodes based on the degree hierarchy, in a label-blind manner (a) and in a differentiated way based on label (b).

chooses nodes from the degree ranking to maximize social influence, setting a threshold $k(n)$ above which all nodes of higher degree are chosen as seeds (Figure 4 (a)).

Hypothesis: Our claim is that color-blind seeding is not necessarily as efficient as color-aware seeding, meaning that choosing only from the highest degree nodes is equivalent to choosing only from the B community in a glass ceiling context, effectively ignoring a whole community, the R nodes. We aim to show that in some conditions choosing differentiated thresholds for the top degree nodes in each community is a better strategy that achieves higher outreach and more diversity simultaneously (Figure 4 (b)). The intuition behind is that at some point there will be more unconverted/uninfluenced nodes in the minority community.

Definition 4.1. The baseline blind strategy defines the seedset as $S_{k(n)} = \{v \in V | deg(v) \geq k(n)\}$.

Definition 4.2. The *influence* of a set S over another set T is defined as the expected number of edges with one end point in S and one in T:

$$\phi_S(T) = |\{(u,v) \in E | u \in S, v \in T\}|. \tag{5}$$

Given this influence, we know that each edge conducts with probability p, so the number of conducting edges within each community is $p \cdot \phi_{S_{k(n)}}(B/R)$.

Goal: Our goal is to find two thresholds $k^R(n)$ and $k^B(n)$ that give in expectation the same amount of seeds as a general ("blind") threshold $k(n)$ but better influence (Figure 4 (b)):

Theorem 4.3. *For the independent cascade model in a network that follows the preferential attachment model with homophily of two communities of red (R) and blue (B) nodes, and a general threshold $k(n)$ for choosing seeds according to their degree, there exists $k^R(n)$ and $k^B(n)$ with $k^R(n) < k^B(n)$, for choosing seeds according to these respective degree thresholds for which*

$$\begin{gathered}\mathbb{E}(\phi(S_{k(n)})) < \mathbb{E}(\phi(S_{k^R(n)} \cup S_{k^B(n)})), \text{ under the constraint} \\ \mathbb{E}(|S_{k(n)}|) = \mathbb{E}(|S_{k^R(n)} \cup S_{k^B(n)}|).\end{gathered} \tag{6}$$

We only give a proof sketch of the above result due to space considerations. Assume that at each timestep t, a new node comes into the network and forms exactly d edges according to the preferential model with homophily, for $d \geq 1$. Then we can write the expected influence a seed set $S_{k(n)}$ has on the network as

$$\begin{gathered}\mathbb{E}(\phi(S_{k(n)})) = \\ n \cdot \mathbb{P}(v \in B) \cdot \mathbb{P}(v \text{ influenced by one of its } d \text{ edges} | v \in B) + \\ n \cdot \mathbb{P}(v \in R) \cdot \mathbb{P}(v \text{ influenced by one of its } d \text{ edges} | v \in R).\end{gathered} \tag{7}$$

Computing these probabilities for the preferential attachment model (omitted here due to space constraints), we get that

$$\begin{aligned}\mathbb{E}(\phi(S_{k(n)})) = n \cdot (1-r) \cdot \Bigg(1 - \Bigg(1 - 2\Bigg(\frac{\frac{1-r}{\alpha\rho+1-\alpha}}{\frac{r\rho}{\alpha+\rho(1-\alpha)} + \frac{1-r}{\alpha\rho+1-\alpha}} \cdot (1-\alpha) \\ \cdot k(n)^{2-\beta(B)} + \frac{\frac{\rho(1-r)}{\alpha\rho+1-\alpha}}{\frac{r}{\alpha+\rho(1-\alpha)} + \frac{\rho(1-r)}{\alpha\rho+1-\alpha}} \cdot \alpha \cdot k(n)^{2-\beta(R)}\Bigg)\Bigg)^d\Bigg) + \\ n \cdot r \cdot \Bigg(1 - \Bigg(1 - 2\Bigg(\frac{\frac{r}{\alpha+\rho(1-\alpha)}}{\frac{r\rho}{\alpha+\rho(1-\alpha)} + \frac{1-r}{\alpha\rho+1-\alpha}} \cdot (1-\alpha) \cdot k(n)^{2-\beta(B)} + \\ \frac{\frac{r}{\alpha+\rho(1-\alpha)}}{\frac{r}{\alpha+\rho(1-\alpha)} + \frac{\rho(1-r)}{\alpha\rho+1-\alpha}} \cdot \alpha \cdot k(n)^{2-\beta(R)}\Bigg)\Bigg)^d\Bigg)\end{aligned} \tag{8}$$

We note that in computing these probabilities in closed form we are limited by the un-constrained nature of the cascading process. For the purpose of this study we make the assumption that the process of diffusion expands to two steps, friends of friends, with the hope of generalizing this work in the future. Since we want to find $k^R(n,x)$ and $k^B(n,x)$ that satisfy our goal in equation 6, we

set $k^B(n,x) = k(n) \cdot x$, for a variable x, and find that solving for the constraint,

$$k^R(n,x) = \left(\frac{1-\alpha}{\alpha} \cdot \frac{\beta(B)-2}{\beta(B)-1} \cdot \frac{\beta(R)-1}{\beta(R)-2} \cdot k(n)^{1-\beta(B)} \cdot (1 - x^{1-\beta(B)}) + k(n)^{1-\beta(R)} \right)^{\frac{1}{1-\beta(R)}}. \quad (9)$$

Then, we can write

$$F(x) = \mathbb{E}(\phi(S_{k^B(n,x)} \cup S_{k^R(n,x)})) - \mathbb{E}(\phi(S_{k(n)})), \quad (10)$$

with the goal of finding x such that $F(x) > 0$. For such an x, it would mean that a more diverse choice of seeds can actually achieve a better outreach, compensating for the intrinsic bias of a network.

We observe that $F(1) = 0$, and if $x < 1$, we are essentially lowering the threshold for the B nodes to make it to the seed set. Since the seed set is already consisting of mostly B nodes, we would like to do the reverse in order to increase diversity, namely to choose $x > 1$ for which $F(x) > 0$. Empirical simulations of this functions show that there are ranges of parameters for which $F(x) > 0$ for $x \in [1, 1+m]$ for some $m \in \mathbb{R}_+^*$. Thus, it is sufficient to show find the ranges of parameters for which $\frac{\partial F}{\partial x}(1) > 0$, which would mean that the function F is increasing around $x = 1$, so it will be positive for a range of $x > 1$ (note: F is continuous in x). The fact that $k^R(n) < k^B(n)$ shows that there is a less stringent threshold for the red nodes, so they will get better represented in the seed selection given their deficiency in the degree distribution.

5 DISCUSSION

These results show the interplay between being fair and strategic: starting with the most basic heuristic, such as the greedy algorithm, being strategic also leads to being fairer. However, large real-world datasets may not always offer the possibility (computationally or feature-wise) to achieve this, and often times we must make use of available metrics, such as degree, distance centrality, and so on, that may be blind to sensitive features. In networks where there exists inequality based on sensitive features, we must ensure that it does not get propagated through our algorithmic design. Our work shows a first step in mitigating the bias that a feature-blind algorithm perpetuates, showing how strategic choices can achieve both diversity in information diffusion and better efficiency. We are able to design such strategic choices with a mathematical guarantee of optimality and fairness.

Future work should analyze the parameter conditions for which strategic seeding as defined above achieves better outreach and diversity in different networks and assess its quantitative benefit. Our method opens up a set of questions regarding the general nature of such trade-offs between fairness and efficiency. While the nature of the independent cascade model favors diversity in maximizing outreach, other models may do the opposite. Indeed, an interesting corollary would be to study critical mass models, such as the linear threshold model. In such cases, a strategic algorithm would focus on achieving that critical mass for information diffusion locally, which may constrain information within one community and prevent it from being equally distributed across the network.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (May 2016).

[2] Chen Avin, Avi Cohen, Pierre Fraigniaud, Zvi Lotker, and David Peleg. 2018. Preferential attachment as a unique equilibrium. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 559–568.

[3] Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. 2015. Homophily and the glass ceiling effect in social networks. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 41–50.

[4] Eric Balkanski, Nicole Immorlica, and Yaron Singer. 2017. The Importance of Communities for Learning to Influence. In *Advances in Neural Information Processing Systems*. 5864–5873.

[5] Solon Barocas and Moritz Hardt. 2017. Fairness in Machine Learning Tutorial. *Neural Information Processing Systems* (2017).

[6] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 199–208.

[7] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[9] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[10] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv preprint arXiv:1610.02413* (2016).

[11] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[12] David Kempe, Jon Kleinberg, and Éva Tardos. 2005. Influential nodes in a diffusion model for social networks. In *International Colloquium on Automata, Languages, and Programming*. Springer, 1127–1138.

[13] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. *arXiv preprint arXiv:1706.02744* (2017).

[14] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[15] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli Upfal. 2000. Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, 57–65.

[16] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *arXiv preprint arXiv:1703.06856* (2017).

[17] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM* 11, 5th (2011), 25.

[18] Lior Seeman and Yaron Singer. 2013. Adaptive seeding in social networks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 459–468.

[19] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 923–932.

[20] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.

[21] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 22.

[22] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

[23] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. PMLR, Atlanta, Georgia, USA, 325–333. http://proceedings.mlr.press/v28/zemel13.html

Hegemony in Social Media and the effect of recommendations

Ana-Andreea Stoica
Columbia University
New York, USA
astoica@cs.columbia.edu

Augustin Chaintreau
Columbia University
New York, USA
augustin@cs.columbia.edu

ABSTRACT

As today's media landscape is carved by social media endorsements and built on automated recommendations, both of these are often criticized for inducing vicious dynamics, such as the filter bubble effect, echo chamber, or polarization. We introduce a new model featuring a mild version of homophily and two well-known popularity dynamics. These broadly reproduce the organic activity and the algorithmic filtering, respectively, of which the latter is now commonplace within social media or other online services. Surprisingly, we show this is all that is needed to create hegemony: a single viewpoint (or side) not only receives undue attention, but it also captures all the attention given to "top trending" items.

CCS CONCEPTS

• **Theory of computation** → **Network flows**; **Social networks**; *Random walks and Markov chains*; • **General and reference** → *Empirical studies.*

KEYWORDS

fairness, recommendation, homophily, random walks

ACM Reference Format:
Ana-Andreea Stoica and Augustin Chaintreau. 2019. Hegemony in Social Media and the effect of recommendations. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317589

1 INTRODUCTION

Item recommendation is instrumental for an online marketplace to develop, for a content provider to grow its audience, and for a social media participant to navigate the deluge of information produced in real time. This is the primary form of personalization, and its effects on growth are beyond doubt: automated recommendations are the origin for 35% of sales on `amazon.com`, 50% of initial messages sent on `match.com`, and 80% of streamed hours on `netflix.com`. Their side effects have never been so scrutinized: concerns range from recommendation algorithms isolating information seekers from differing viewpoints (*filter bubble*), radicalizing citizens' attitude towards controversial issues (*polarization*), or enabling malicious actors to manipulate information spreading (*fake news*). It is increasingly common that opting out of recommendations is simply not an option: accessing news on Facebook or searching for partners on Tinder. Still, many users express a concern that the results they see may not be representative.

We describe a model to study whether diverse viewpoints are rewarded with attention from social media and recommender systems. The alternative would be a form of *hegemony* where, although multiple viewpoints exist on a controversial topic, only one of them effectively reaches a large audience. Social media are often criticized for their alleged effects on opinion formation. They claim a form of neutrality: that they add no supplementary bias to the relative proportion of those supporting different viewpoints. In particular, they vehemently deny that item recommendations or the selection of *top trending* items leverages any feature or design that could favor a particular ideology. That claim seems *a priori* credible since all media know that news readers massively oppose partisanship and ideological bias, with up to 75% claiming it is "never acceptable"[13]. But the paradox is, despite all of the above, there is a considerable distrust in the media, especially social media and algorithmic recommendations, to represent each viewpoint fairly. This lack of trust has dangerous consequences and often accompanies accusations of dishonesty. Here we explore an alternative explanation where this chasm comes from a natural hegemony emerging from root causes commonly found in most online services.

We assume no pre-existing social network connecting readers that could affect their content exposure[1]. We assume a mild form of homophily, where users tend to connect more to content they agree with. Under these (idealized) conditions, we answer the following questions: does the organic dynamics of posting and reposting in social media offer an outlet for all viewpoints to be shown? Can one viewpoint become hegemonic? And is a simple recommendation algorithm such "people who like ... also like" a deciding or a contributing factor?

- We first analyze a simple content-linking model. Readers implicitly belong to one of two groups (e.g. liberal or conservative, female or male). We prove that social media reposting *systematically under-represent* a minority viewpoint, the majority viewpoint gradually becoming the *only* visible option. (Section 4)
- We extend our model to more complex forms of homophily, and show it accurately reproduces an observed gap in the amount of likes received by items on a leading social media (i.e. pictures on Instagram). Interestingly, it shows that hegemony may be created not by majority alone but also

This material is based upon work supported by the National Science Foundation under Grant No.1254035 and 1514437.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317589

[1]Note that exposure to ideologically diverse news on Facebook was studied empirically and shown to present a bias that was primarily attributed to that effect [3].

by minority when it exhibits exacerbated homophily. (Section 5).
- Finally, we study the effect of recommendations. We find that items recommendations *always accelerate* the hegemonic dynamics we identified, exacerbating the misrepresentation of a minority viewpoint. (Section 6).

2 RELATED WORK

In spite of a large body of work studying the effect of recommendation systems, we found no model that reproduces the simple underlying dynamics of a hegemonic nature we observe. A critical novelty of our approach is that we study the effect of recommendations within the natural organic growth of the social media. Empirically, recommendations were shown to influence users in creating blockbusters [14], to expose users to a slightly narrowing set of items [10], or on the contrary to widen their interests [8]. A model of recommendation based on market share was shown to lead to higher concentration overall, while individuals enjoy higher product diversity [6]. We study dynamics that more generally emerge from the social endorsements created with time and exploited to build recommendations.

Filter bubble [12], balkanization [16] or polarization [4, 17] dynamics have been suspected to play an increasing role online for almost a decade. Polarization and echo-chambers were identified before in pre-existing structures among blogs or in the social network connecting individuals [1, 3], with important consequences on information exposure. However, recent evidence relativizes those hypotheses [5] as it finds only a modest segregation or ideological distance between news readers. Our analysis points to a different hypothesis: first, our model has no graph connecting participants that would limit or direct their information exposure; second, we do not predict that two groups of readers online end up at polar opposites, but rather explain why dissenting and minority views are suffering from a lack of representation. We note that, in contrast with previous work, these dynamics emerge in conditions that are almost always met online, which may explain why individuals holding viewpoints that are not shared by the majority found that items recommended in social networks rarely reflect their views. Note that a lack of equal access to representation in today's online services (including Twitter [11], TaskRabbit [7]) was empirically observed. This adds to a large body of work warning us on the unintended consequences of algorithms run on Big Data, with discrimination taking place in seemingly neutral settings. Since the relative representation of various viewpoints is driven by participants' homophilic behavior rather than an algorithm itself, it remains difficult to determine if that constitutes a case of active discrimination. It is however undeniably generating a disparate impact, and the fact that an algorithm reinforces this effect makes a strong case that more work is needed to prevent this harm.

The result most relevant to ours is a recent proof that rich-get-richer dynamics combined with homophily naturally exacerbate the advantage of a majority group [2]. That was recently extended to understand how social recommender system contributes to create a glass-ceiling [15], an effect that is equivalent to hegemony in a unipartite graph.

3 MODELS AND DEFINITIONS

We introduce two models - one describing the organic growth of a network and one in which recommendations are added. Both models extend previous work on network growth [2, 9, 15] by separating nodes into *individuals* and *items* in a bipartite manner.

Participants interact with items on social media by either posting their own content, or by reposting: liking, retweeting, or sharing. For simplicity we assume participants have one color (Red or Blue) that represents their opinion. Items also have colors, inherited from the person who posted it originally. Colors may represent for instance liberal and conservative views on an issue, but the model is general and can be applied to goods bought online from different brands, job ads, etc.

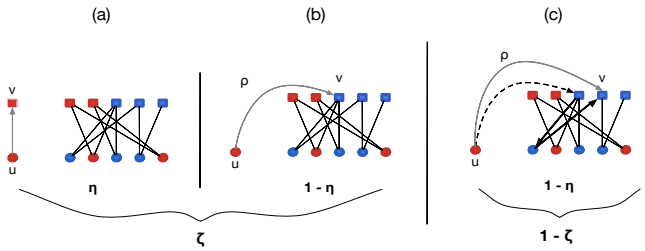


Figure 1: Illustration of the organic growth model $\mathcal{M}$ and the recommendation model $\mathcal{M} - REC$, where individuals are circles and items are squares.

3.1 Organic growth model $\mathcal{M}$

We introduce a model of growth $\mathcal{M}$, inspired by the biased preferential attachment model [2]:

- *Minority-majority partition*: for $0 \leq r \leq 1/2$, at time t, an individual enters the network and receives label R with probability r and B with probability $1 - r$.
- *Item creation*: with probability η, the new individual creates a new item that has the same color as herself, and connects to it with probability 1. This step allows the set of items to grow, capturing the dynamics or creativity in a social network - for example, an author of a certain political view writes and publishes an article that reflects her view, or composes a tweet (Figure 1 (a)).
- *Reposting*: with probability $1 - \eta$, the new individual reposts/retweets/likes one existing item. Intuitively, items that have been reposted more are more likely to be noticed and hence chosen to be reposted. We thus assume the individual chooses one item to connect to with probability proportional to that item's degree, $\mathbb{P}(v \text{ is chosen}) = \delta_t(v) / \sum_{u \in V_t} \delta_t(u)$, where $\delta_t(x)$ denotes the degree of item x at time t and V_t is the set of items in the graph at time t (Figure 1 (b)).
- *Homophily:* if the individual has a different label than the item it chooses to connect to in the previous step, the connection is accepted with probability ρ (and rejected with probability $1 - \rho$). If rejected, the process is repeated until an edge is formed. The homophily parameter $0 \leq \rho \leq 1$ captures a person's openness to repost content that does

not match her views; $\rho = 1$ denotes an equanimous person which is not affected by that item's viewpoint in any way. We typically assume that $\rho < 1$. Intuitively, most individuals have a tendency to read broadly across the political spectrum, but are more selective in the ideology of items they repost [1].

3.2 Recommendation model $\mathcal{M} - REC$

We extend the model $\mathcal{M}$ by adding recommendations based on individuals with similar behavior. This captures the way in which, for example, an individual would see a link posted by a friend with whom he shared previous items.

We denote this new model by $\mathcal{M} - REC$ and define it below:

- *Minority-majority partition*: for $0 \leq r \leq 1/2$, at time t, an individual enters the network and receives label R with probability r and B with probability $1 - r$.
- with probability ζ, it connects to an item according to model $\mathcal{M}$ (Figure 1 (a) and (b)).
- with probability $1 - \zeta$, it chooses an item according to preferential attachment, but it does not connect to it; instead, it follows a 3-step random walk from that item by finding someone who was connected to that item and choosing another item linked to that individual. If the end item has the same color, it connects to it, and if not, it accepts it with probability ρ and rejects with probability $1 - \rho$, and repeats the process until an edge is formed (Figure 1 (c)).

Note: We assume each node has an outdegree of $m > 1$ and exclude it from the analysis it will just create a m-factor multiplication.

We adapt the analysis of the glass ceiling effect in the unipartite graph from [2], where two communities exhibit a glass-ceiling effect if their degree distributions follow a power law with different coefficients. Intuitively, this shows that the gap between them would increase as we are looking at higher-ranked individuals.

4 HEGEMONY IN ORGANIC GROWTH

We first present a theoretical result, where the item-based preferential attachment model $\mathcal{M}$ exhibits a hegemony effect, effectively suppressing the representation of the items pertaining to the minority viewpoint in top results.

A similar proof exists for a more constrained model on unipartite graph [15]. An important step is to obtain the fraction of connections to red-items as a fixed point equation. We find that, provided the fixed point equation is adapted to this new case, similar arguments can be used for a bipartite graph.

Rate at which red items receive connections: Denote by $u_t(R)$ as the sum of degrees of all red items after t steps. Since an edge is added at each step, it makes sense to introduce $u_t(R) = t \cdot \alpha_t$ and $u_t(B) = t \cdot (1 - \alpha_t)$, where α_t is equivalently defined as the sum of degree of the red items divided by the total sum of their degrees, or simply the fraction of edges created towards R items.

Lemma 4.1. $\mathbb{E}[\alpha_{t+1}|\alpha_t] = \alpha_t + \frac{F(\alpha_t)-\alpha_t}{t+1}$, *where*

$$F(x) = \eta \cdot r + (1-\eta) \cdot \left(\frac{rx}{x + \rho \cdot (1-x)} + \frac{x \cdot \rho \cdot (1-r)}{x\rho + 1 - x} \right).$$

We observe similar properties of the F function than the other function described for unipartite graph in Avin et al [2]. For space considerations, we omit the complete argument of this proof.

Lemma 4.2. *For $r < 1/2$ and $0 < \rho < 1$, we have*

(1) *F is monotonically increasing.*
(2) *F has exactly one fixed point in $[0, 1]$, denoted by α^*.*
(3) *$x < \alpha^* \implies x < F(x) < \alpha^*; x > \alpha^* \implies \alpha^* < F(x) < x$.*
(4) *$\alpha^* < r$.*

As Lemma 4.1 illustrates, F is the expected rate of growth of red edges from time t to time $t + 1$. Thus, its fixed point α can be interpreted as the limit of the fraction of red edges when the network size grows to infinity. Such a limit exists since the fixed point is present and unique in $[0, 1]$.

The last property, $\alpha^* < r$, already shows that the expected degree of a red item (α^*/r) is lower than that of a random item (1). However, this does not necessarily imply a hegemonic effect. We now show that the degree distribution of the two types of items follow power laws with different coefficients, which implies hegemony as defined above.

Degree distribution: Denote by $m_{k,t}(B)$ and $m_{k,t}(R)$ the number of blue and red items of degree k at time t, and define

$$M_k(x) = \lim_{t \to \infty} \frac{\mathbb{E}(m_{k,t}(x))}{t}, \text{ for } x \in \{R, B\}. \tag{1}$$

Theorem 4.3. *In the case of organic growth according to model $\mathcal{M}$, the degree distributions of the red and blue items follow power laws, i.e. $M_k(R) \sim k^{-\beta(R)}$ and $M_k(B) \sim k^{-\beta(B)}$.*

We omit the proof, noting that in order to show this theorem we derive a recurrence relation for $m_{k,t}(R)$, $m_{k,t}(B)$ and deduce that degrees quickly converge to a power law as t gets large with coefficients $\beta(R) = 1 + \frac{1}{(1-\eta)C_R}$ and $\beta(B) = 1 + \frac{1}{(1-\eta)C_B}$, where

$$\begin{aligned} C_B &= \frac{\rho \cdot r}{\alpha + \rho(1-\alpha)} + \frac{1-r}{\rho \cdot \alpha + 1 - \alpha}, \\ C_R &= \frac{r}{\alpha + \rho(1-\alpha)} + \frac{\rho \cdot (1-r)}{\rho \cdot \alpha + 1 - \alpha}. \end{aligned} \tag{2}$$

Using the properties of the function F and these results, we find the following corollary:

Corollary 4.4. *For $r \in [0, 1/2]$ and $\rho \in [0, 1]$, we have*

$$\beta(R) > 2 > \beta(B). \tag{3}$$

Gap coefficient and normalized expected degree of a red node: Our work enables us to make a qualitative analysis of how the expected degree of a red node changes when homophily varies between 0 and 1. Firstly, notice that since α represents the expected fraction of edges towards red items, α/r is the normalized expected degree of a red note as the fraction of red items is r.

Furthermore, we are able to compute the gap between the rate at which the number of items above degree k decreases as a function of k. To do this, we are interested in computing the fraction of items with degree at least k: $\mathbb{P}(u \in R | deg(u) > k)$. We make use of the

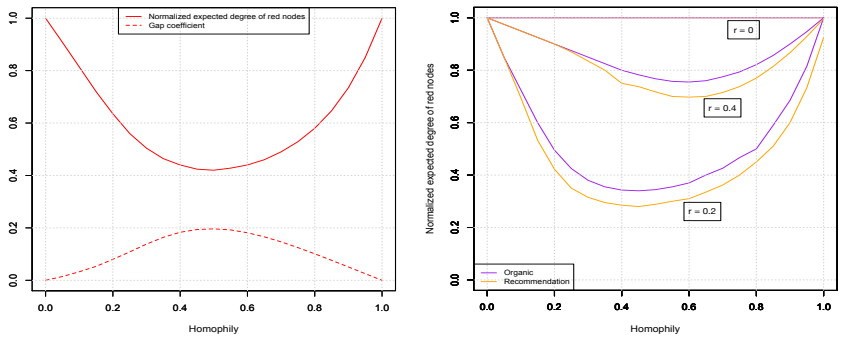


Figure 2: Effect of homophily parameter ρ as x-axis on the normalized expected degree of red nodes and the gap coefficient for $r = 0.25$ and $\eta = 0.1$ (a) and on the normalized expected degree of red nodes, for $r = 0, 0.2, 0.4, \eta = 0.1, \zeta = 0.9$ (b).

fact that the degree distribution of items of degree exactly k follows a power law with known coefficients:

$$\begin{aligned}\mathbb{P}(u \in R, deg(u) > k) &\sim C \cdot k^{-(1+\beta(R))} = C \cdot k^{-(2+1/(1-\eta)C_R)},\\ \mathbb{P}(u \in B, deg(u) > k) &\sim C \cdot k^{-(1+\beta(B))} = C' \cdot k^{-(2+1/(1-\eta)C_B)},\end{aligned} \tag{4}$$

for constants C, C'. Then, for some constants C_1, C_2,

$$\begin{aligned}\mathbb{P}(u \in R | deg(u) > k) &= \frac{\mathbb{P}(u \in R, deg(u) > k)}{\mathbb{P}(deg(u) > k)}\\ = \frac{r \cdot C_1 \cdot k^{-(2+1/(1-\eta)C_R)}}{r \cdot C_1 \cdot k^{-(2+1/(1\eta)C_R)} + (1-r) \cdot C_2 \cdot k^{-(2+1/(1-\eta)C_B)}} &\Rightarrow\\ \mathbb{P}(u \in R | deg(u) > k) &\approx k^{-\frac{1}{1-\eta}\left(\frac{1}{C_R} - \frac{1}{C_B}\right)}.\end{aligned} \tag{5}$$

We call $\frac{1}{C_R} - \frac{1}{C_B}$ the *gap coefficient*. Figure 2 thus shows how the normalized expected degree of a red node (straight red line) and the gap coefficient reach a minimum and a maximum, respectively, for an intermediate value of ρ. Thus, there is a critical value of homophily for which the red items lose the most of their edges.

An example of the effect this can have in the real world is to consider Youtube videos and analyze their revenue as a function of their number of likes. Given Figure 2, one could expect that homophily can reduce the revenue for videos pertaining to a minority view by up to 60% of what they should receive in an equal setting.

5 VALIDATION AND EXTENSIONS

5.1 Empirical study of Hegemony

We validate our results using data collected from Instagram to study the effect of homophily among gender on the popularity of items. To the best of our knowledge, no prior studies of this case exists. The data will be made publicly available. For the purpose of this study, we use a subset of this dataset, consisting of 92, 935 users with labeled gender, who posted in total 44, 725, 839 photos. We consider each of these photo as an "item", and we attribute it the gender of its author.

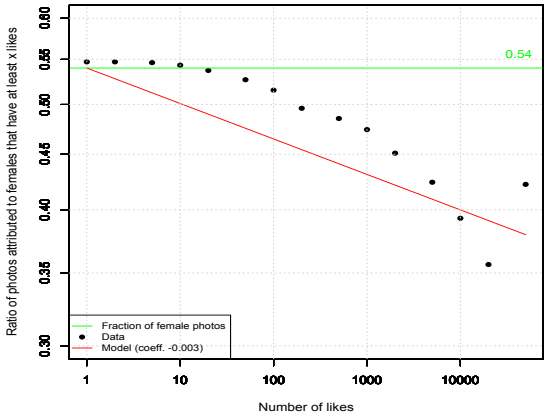


Figure 3: Fraction of photos posted by females among those that have at least x likes for the Instagram dataset (black dots) and model $\mathcal{M}$ (red line), where the power law coefficient is estimated to be a= -0.0325, plotted in log-log scale. The green line is the true fraction of female photos, i.e. 54%.

Figure 3 presents the observed fraction of female photos among those with at least k likes. As k grows, we focus on those that are most likely to be noticed and feature in a "top trending" set of items. Unfortunately, that seems to have a detrimental effect on the observed fraction of female pictures, which drops from 54% (the fraction of female accounts) to 40%. We note that the trend, mild at first, seems to accelerate towards the top.

At first, this seems the exact opposite of our model, which predicts that female should be more represented since they are a majority. In order to reconcile those results with this empirical analysis, we extend the present theoretical framework by introducing *differentiated homophily*, case in which the two subpopulations exhibit different homophily parameters.

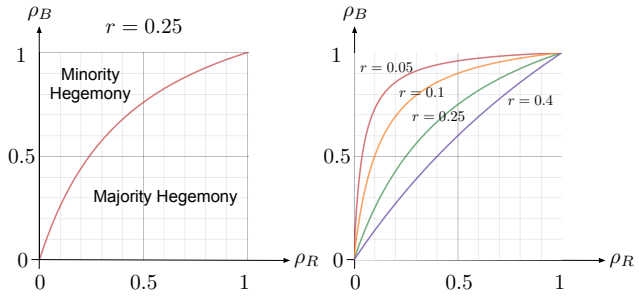


Figure 4: The two regions of hegemony appearing for different values of differentiated homophily ρ_B and ρ_R, plotted here for $r = 0.25$ (left), and various values of r (right).

5.2 Extension to differentiated homophily

As our main result, we find that hegemony is the norm, rather than the exception, for a more general model of homophily.

Theorem 5.1. *If model $\mathcal{M}$ exhibits differentiated homophily ρ_R and ρ_B, in which a red node rejects a blue node with probability ρ_R and a blue node rejects a red node with probability ρ_B, then the network exhibits:*

- *A hegemonic majority iff* $(1-r)\cdot\left(\frac{1}{\rho_B}-1\right) > r\cdot\left(\frac{1}{\rho_R}-1\right)$.
- *A hegemonic minority iff* $(1-r)\cdot\left(\frac{1}{\rho_B}-1\right) < r\cdot\left(\frac{1}{\rho_R}-1\right)$.
- *No hegemony iff* $(1-r)\cdot\left(\frac{1}{\rho_B}-1\right) = r\cdot\left(\frac{1}{\rho_R}-1\right)$.

The proof of this theorem leverages new properties of the function F to show that the above comparison determines whether $F(r) < r$ holds. This theorem confirms the previous empirical observations that, in certain conditions, a minority with a stronger homophily may overcome its disadvantage and obtain an hegemonic advantage. We present in Figure 4 the various regions of (ρ_R, ρ_B) under which the two types of hegemony occurs. Note that majority hegemony occurs across a larger range of parameters.

We empirically observed that female users on Instagram are less homophilic: they are more likely to like a picture of a male than vice-versa. The homophily parameter is hence smaller for males than for females ($\rho_R < \rho_B$) which would be consistent with Theorem 5.1.

6 HEGEMONY UNDER RECOMMENDATIONS

In the extended model of recommendation, computing the evolution equation for the degree of the items becomes more complicated since we include the case of the random walk of length 3. Thus, we make a couple of assumptions for ease of work:

- edges in the random walk are assumed to have formed through model $\mathcal{M}$.
- edges in the random walk are independent of each other.
- edges that already exist are naturally formed at different steps in time; however, when computing an evolution equation, we set their time of formation to the current time, t.

We first start by investigating again the equilibrium state when the number of red edges converges as the fixed point of a function that describes their growth.

Rate at which red items receive connections: Denote by X_0, X_1, X_2, X_3 the vertices of the random walk created through recommendation, where X_0 is the newly added node to the network (thus, X_0 and X_2 are individuals and X_1 and X_3 are items).

Denoting again by R_{t+1} the number of red balls at time t and by $\alpha_{2,t}$ the fraction of red balls at time t, we observe the following cases in computing $\mathbb{P}(R_{t+1} = 1)$:

- a red item obtain a connection through model $M2$, with probability

$$\zeta\cdot\left(\eta\cdot r + (1-\eta)\cdot\left(\frac{r\cdot\alpha_{2,t}}{\alpha_{2,t}+\rho(1-\alpha_{2,t})} + \frac{(1-r)\cdot\rho\alpha_{2,t}}{\rho\alpha_{2,t}+1-\alpha_{2,t}}\right)\right). \quad (6)$$

- a red item obtain a connection through recommendation, with probability

$$(1-\zeta)\cdot\left(\mathbb{P}(X_0\in R)\frac{\mathcal{P}(X_3\in R|X_0\in R)}{\mathcal{P}(X_3\in R|X_0\in R)+\rho\cdot\mathcal{P}(X_3\in B|X_0\in R)} + \mathbb{P}(X_0\in B)\frac{\rho\cdot\mathcal{P}(X_3\in R|X_0\in B)}{\mathcal{P}(X_3\in B|X_0\in B)+\rho\cdot\mathcal{P}(X_3\in R|X_0\in B)}\right) \quad (7)$$

For ease of computation, consider the probabilities from above as functions of $\alpha_{2,t}$ (so for example, $\mathcal{P}(X_3\in R|X_0\in R) = \mathcal{P}(X_3\in R|X_0\in R)(\alpha_{2,t})$). Thus, we define a new function F_2, similar as for the organic growth model, which takes the following form (for detailed computation of the probabilities, refer to appendix A):

$$F_2(x) = \zeta\cdot F(x) + (1-\zeta)\cdot\left(r\cdot\frac{\mathbb{P}(X_3\in R|X_0\in R)(x)}{\mathbb{P}(X_3\in R|X_0\in R)(x)+\rho\cdot\mathbb{P}(X_3\in B|X_0\in R)(x)} + (1-r)\cdot\frac{\rho\cdot\mathbb{P}(X_3\in R|X_0\in B)(x)}{\mathbb{P}(X_3\in B|X_0\in B)(x)+\rho\cdot\mathbb{P}(X_3\in R|X_0\in B)(x)}\right), \quad (8)$$

where $F(x)$ is the function from model $\mathcal{M}$.

Computing each of these probabilities in closed form, we are able to show that F_2 exhibits the exact same properties as the function F (Lemma 4.2), having a fixed point α_2^*. This shows that the fraction of red edges also converges towards a constant that depends on r, ρ, η, ζ, as it is a stable point at the intersection of $F_2(x)$ and $f(x) = x$. Most importantly, empirical analysis reveals that $\alpha_2 < \alpha$ in this case as well, showing that the red items have less power under recommendation than under organic growth.

We continue by showing that the degree distribution also follows a power law and by showing that the coefficients are even further apart than in organic growth.

Degree distribution: Denote by $m^2_{k,t}(B)$ and $m^2_{k,t}(R)$ the number of blue and red items of degree k at time t in the model $\mathcal{M}-REC$, and define

$$M^2_k(x) = \lim_{t\to\infty}\frac{\mathbb{E}(m^2_{k,t}(x))}{t}, \text{ for } x\in\{R,B\}. \quad (9)$$

Theorem 6.1. *In the case of recommendation growth according to model $\mathcal{M}-REC$, the degree distribution of the red and blue items also follows a power law, i.e. $M^2_k(R)\sim k^{-\beta_2(R)}$ and $M^2_k(B)\sim k^{-\beta_2(B)}$ for coefficients $\beta_2(R)$ and $\beta_2(B)$.*

We prove this by deriving again a recurrence relation for $m^2_{k,t}(R)$ and $m^2_{k,t}(B)$ (omitted here) and computing the coefficients as

$$\beta_2(R) = 1 + \frac{1}{(1-\eta)C_{2,R}} \text{ and} \beta_2(B) = 1 + \frac{1}{(1-\eta)C_{2,B}},$$

where $C_{2,R}$ and $C_{2,B}$ can again be computed explicitly.

Theorem 6.2. *For $r \in [0, 1/2]$ and $\rho \in [0, 1]$, we have*

$$\beta_2(R) > \beta(R) > 3 > \beta(B) > \beta_2(B), \tag{10}$$

implying again that the tail and strong glass ceiling effects are even more pronounced under algorithmic recommendations (model $\mathcal{M}-REC$) than under organic growth (model $\mathcal{M}$).

Proof. In proving this, we derive an invariant equation for the rate of growth of red edges and use that $\alpha_2^* < \alpha^*$. □

Lemma 6.3. *Given α and α_2 the limits of the fraction of red balls as defined above, the following holds:*

$$(1-\eta)\cdot\alpha\cdot C_R + \eta\cdot r = \alpha, \text{ and } (1-\eta)\cdot\alpha_2\cdot C_{2,R} + \eta\cdot r = \alpha_2. \tag{11}$$

Proof. As C_R encapsulates the rate at which red edges appear, the system reaches an equilibrium state. Due to this, the rate at which red edges appear must equal the current fraction of red edges, as it does not evolve anymore. Furthermore, as η represented the probability for an item to be created, $\eta \cdot r$ is the rate of growth of edges towards red items in the case where they are created, and $(1-\eta)\cdot\alpha\cdot C_R$ is the rate of growth of edges towards red items in the preferential attachment case. The sum of these is then equal to α as this is the number of red edges at equilibrium. For the second part of equation 11, note that while ζ plays a role, our assumption is that the edges form sparsely through recommendation, so we may approximate the rate of growth of red edges at equilibrium. In a sense, since this result holds even when "ignoring" such edges formed through recommendation in this invariant equation, we conclude that their influence is even higher in reality. □

Subtracting the equations in 11, we obtain

$$\begin{aligned} (1-\eta)\cdot(\alpha\cdot C_R - \alpha\cdot C_{2,R}) &= \alpha - \alpha_2 \Leftrightarrow \\ (1-\eta)\cdot(C_R\cdot(\alpha-\alpha_2) + \alpha_2\cdot(C_R - C_{2,R})) &= \alpha - \alpha_2 \Leftrightarrow \\ (1-\eta)\cdot\alpha_2\cdot(C_R - C_{2,R}) &= (\alpha-\alpha_2)\cdot(1-(1-\eta)\cdot C_R). \end{aligned} \tag{12}$$

As shown before, $C_R < 1$ and $\alpha_2 < \alpha$, which yields $C_R > C_{2,R}$, since $\eta < 1$. Since $\beta(R) = 1 + \frac{1}{(1-\eta)C_R}$ and $\beta_2(R) = 1 + \frac{1}{(1-\eta)C_{2,R}}$, Theorem 6.2 is proved. The case for C_B and $C_{2,B}$ can be proved similarly. Thus, we show that the tail and moment hegemony are even more exacerbated in item recommendations than in the organic network.

Normalized expected degree of red nodes: We analyze comparatively the effect of homophily on the expected degree of red nodes for the organic growth model $\mathcal{M}$ and the recommendation model $\mathcal{M}-REC$, for different values of r. As compared to Figure 2 (a), we observe in Figure 2 (b) that recommendations exacerbate the gap between the two subgroups. Computing the normalized degree of red nodes as α/r (purple line) and α_2/r (orange line), recommendations decrease the expected degree of red items by 5% for both $r = 0.2$ and $r = 0.4$.

7 CONCLUSION

Our theoretical analysis unveils the subtle relation between network structure, homophily, and hegemony. By exploring a model of growth based on item recommendations, we show that in the case of a bi-populated network, the minority's viewpoint loses representation. It is interesting to notice that the minority viewpoint is not excluded from the network, as the ratio of items pertaining to the minority remains constant throughout, but it is only at the top of the hierarchy where this effect occurs. This effect has great ramifications in the way we interact with social media and the information that is disseminated through popular items. Future work should focus on extended empirical analysis and on the design of algorithms that alleviate such effects by taking into consideration network structure.

REFERENCES

[1] J An, D Quercia, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2014. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science* (2014).

[2] C Avin, B Keller, Z Lotker, C Mathieu, and David Peleg. 2015. Homophily and the glass ceiling effect in social networks. *ITCS* (2015).

[3] E. Bakshy, S Messing, and L.A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[4] Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (April 2013), 5791–5796.

[5] S Flaxman, S Goel, and J M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quaterly* 80, 1 (2016), 298–320.

[6] D Fleder and K Hosanagar. 2009. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* 55, 5 (May 2009), 697–712.

[7] Aniko Hannak, C Wagner, David Garcia, Markus Strohmaier, and Christo Wilson. 2016. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit. *DAT '16: Proceedings of The Workshop on Data and Algorithmic Transparency* (Nov. 2016), 1–8.

[8] K Hosanagar, D Fleder, D Lee Management Science, and 2013. 2008. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Science* 60, 4 (2008), 805–823.

[9] Ravi Kumar, Prabhakar Raghavan, S Rajagopalan, D Sivakumar, Andrew Tomkins, and E Upfal. 2000. Stochastic models for the Web graph. (2000), 57–65.

[10] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW '14: Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee.

[11] S Nilizadeh, A Groggel, P Lista, and S Das. 2016. Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility. *Proceedings of the International Conference Weblogs and Social Media (ICWSM)* (2016).

[12] E Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin.

[13] K Purcell, L Rainie, A Mitchell, T Rosenstiel, and K Olmstead. 2010. Understanding the participatory news consumer. *Pew Research Center's for Excellence in Journalism* (2010).

[14] M J Salganik, P S Dodds, DJ Watts science, and 2006. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311 (2006).

[15] Ana Stoica, Christopher J Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks. *WWW '18: Proceeding of the 25th International Conference on World Wide Web* (2018), 1–9.

[16] Marshall Van Alstyne and Erik Brynjolfsson. 2005. Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities. *Management Science* 51, 6 (June 2005), 851–868.

[17] Ingmar Weber Venkata Rama Kiran Garimella. 2017. A Long-Term Analysis of Polarization on Twitter. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media ICWSM* (April 2017), 1–4.

Privacy and Transparency within the 4IR: Two faces of the same coin

Bianca Rodrigues Teixeira

Department of Informatics, Pontifical Catholic University, Rio de Janeiro, Brazil
bteixeira@inf.puc-rio.br

Daniel Schwabe

Department of Informatics, Pontifical Catholic University, Rio de Janeiro, Brazil
dschwabe@inf.puc-rio.br

Fernanda A. Baião

Department of Informatics, Pontifical Catholic University, Rio de Janeiro, Brazil
fbaiao@inf.puc-rio.br

Flávia M. Santoro

Postgraduate Program in Computational Sciences, University of the State of Rio de Janeiro, Rio de Janeiro, Brazil
flavia@ime.uerj.br

Maria Luiza M. Campos

Department of Computer Science, Federal University or Rio de Janeiro, Rio de Janeiro, Brazil
mluiza@ppgi.ufrj.br

Leticia D. Verona

Postgraduate Program in Informatics, Federal University or Rio de Janeiro, Rio de Janeiro, Brazil
leticiaverona@ufrj.br

Carlos Laufer

Department of Informatics, Pontifical Catholic University, Rio de Janeiro, Brazil
laufer@globo.com

Simone Barbosa

Department of Informatics, Pontifical Catholic University, Rio de Janeiro, Brazil
simone@inf.puc-rio.br

Sergio Lifschitz

Department of Informatics, Pontifical Catholic University, Rio de Janeiro, Brazil
sergio@inf.puc-rio.br

Rosa Maria E. Moreira

Postgraduate Program in Computational Sciences, University of the State of Rio de Janeiro, Rio de Janeiro, Brazil
rcosta@ime.uerj.br

ABSTRACT

The Fourth Industrial Revolution (4IR) is characterized by a fusion of technologies, which is blurring the lines between the physical, digital, and biological spheres. In this context, two fundamental characteristics emerge: transparency and privacy. From one side, transparency can be seen as the quality that allows participants of a community to know which particular processes are being applied, by which agents, and on which data items. It is generally regarded as a means to enable checks and balances within this community, so as to provide a basis for trust among its participants. Privacy, on the other side, essentially refers to the right of an individual to control how information about her/him is used by others. The issue of public transparency versus individual privacy has long been discussed, and within already existing 4IR scenarios, it became clear that the free flow of information fostered by transparency efforts poses serious conflicting issues to privacy assurance. In order to deal with the myriad of often conflicting cross-cutting concerns,

Internet applications and systems must incorporate adequate mechanisms to ensure compliance of both ethical and legal principles. In this paper, we use the OurPrivacy Framework as a conceptual framework to precisely characterize where in the design process the decisions must be made to handle both transparency and privacy concerns.

CCS CONCEPTS

• Information systems → Social networks • **Information systems → Semantic web description languages**
• Security and privacy → Social aspects of security and privacy

KEYWORDS

Privacy, Transparency, Trust, Semantic Web, Policy, Nanopublications, Knowledge Graph.

1 Introduction

The years of the Third Industrial Revolution (3IR), also called the Digital Age, when digital technologies enabled new ways of generating, processing and sharing information, are ending. According to Klaus Schwab, the founder of the World Economic Forum, we are moving into the Fourth Industrial Revolution (4IR) [12], which builds on 3IR and adds a fusion of technologies, thus blurring the lines between the physical, digital, and biological spheres.

As the daily lives of billions of people are affected – and dependent – on the flow of information (and, ultimately, on knowledge) [14], two important characteristics of the use of such information emerge: transparency and privacy.

Transparency can be seen as the quality that allows participants of a community to know which particular processes are being applied, by which agents, and on which data items. It is generally regarded as a means to enable checks and balances within this community, so as to provide a basis for trust among its participants. If we take the whole society as a community, these checks and balances are reflected on its political system, to prevent misuse by any of the parties involved. Particularly, one of the mechanisms created to increase Transparency in political systems is the enactment of legislation ensuring the right of its members (i.e., citizens in general) to access, create or publish data from a variety of contexts, ranging from the details and results of government acts to consumer-related details of goods and products, as well as the right of individuals to freely create, publish and access information [10].

Privacy, on the other hand, is a basic human right [27], but because of the lack of consensus on its concept, it is difficult to define what it means. Currently, privacy is a comprehensive concept, encompassing freedom of thought, control over the body, isolation, control over personal information, freedom of vigilance, reputation protection and protection against searches and investigations [7]. Theorists have characterized privacy in terms of access [28], or the merit that gives an individual the ability to control access that others have to him/her [1]. Privacy is also defined as the ability to determine when, how, and to what extent information about us is communicated to other people [2]. Despite its several definitions in the literature, it essentially refers to the right of an individual to control how others use information about her/him.

The issue of public transparency versus individual privacy has been long discussed, most particularly Aristotle's distinction between the public sphere of political activity and the private sphere of domestic life. Within already existing 4IR scenarios, it became clear that the free flow of information fostered by transparency efforts poses serious conflicting issues to privacy assurance. In order to deal with the myriad of often conflicting cross-cutting concerns, Internet applications and systems must incorporate adequate mechanisms to ensure compliance of both ethical and legal principles.

In this paper, we use OurPrivacy [5] as a conceptual framework of reference to precisely characterize where in the design process the decisions must be made by designers to handle both transparency and privacy concerns, and how it can support accountability.

2 Background on Transparency and Privacy

According to Meijer [3] and others [14][29][16][19], transparency can be defined as "the availability of information about an actor that allows other actors to monitor the workings or performance of the first actor." It contemplates the "capacity of outsiders to obtain valid and timely information about the activities of government or private organizations" [18].

As a relational concept, transparency presupposes the involvement of an observed and an observer [8]. Besides, there is a difference as to whether the actor that makes information available is an organization or an individual (who may or may not be part of a certain organization) [17] and whether the information relates to an individual or to other more complex events or situations.

Transparency, then, is closely coupled to accountability [25]. However, while transparency is one ingredient in accountability, it is not sufficient. According to Boven [15], accountability essentially means an obligation to explain and justify the conduct of and agent to a third party (the actor and the forum, respectively, in Boven's terminology). Transparency initiatives should provide citizens and other stakeholders with information on the actions of public organizations in order to hold elected officials and public agencies accountable for their decisions and actions. The conceptual complexity of the data openness in the public sector, as opposed to the private sector, relates to several aspects: i) the changing notions of public/private interests; ii) the delicate balance between freedom of information and privacy; iii) government security issues; and iv) the notion that information maintained by the governments is a public asset and should be accountable [24].

Transparency and accountability are also related to privacy. In current business contexts, for example, users/consumers' information plays a central role, supporting sales and marketing strategies, and, most often, strongly influencing and directing consumer behavior. Privacy, in this case, refers to the right of the individual to control how information about him/her is used by others, such as a company. Transparency, on the other hand, supports the enforcement of privacy and other data protection regulations, since the disclosure of information about the company's processes and procedures associated to this individual contributes to trusting that it is indeed compliant with these regulations [17]. As stated by Gutwirth and de Hert [23], data protection and criminal procedures "can be mainly – not exclusively – seen as 'tools of transparency' (regulating and channeling necessary/ reasonable/ legitimate power)" in comparison with privacy, which "is an example of a 'tool of opacity' (stopping power, setting normative limits to power)".

Regulations are important instruments to guarantee rights and duties among involved parties. In general, users' contracts, inter-organizational agreements, domain specific norms and legislation are constantly evolving, reflecting on demands from a changing society, as well as requirements from the adoption of new technologies. The Data Protection Directive of 1995 made no mention to individual rights to data protection, focusing, instead, mostly on controllers' obligations. More recently, the General Data Protection Regulation (GDPR) [21] focuses on "protecting fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data". However, the European Union already had previous regulations concerning privacy, more specifically in the European Convention on Human Rights, which stated the principle that data should only be used for the limited purpose for which it was gathered.

As a data protection regulation, the GDPR imposes greater obligations on actors, taking rights much further [30]. In this context, transparency is explicitly associated to the fact that any request for consent should be intelligible, accessible and expressed in clear language, and forcing disclosure of information about data transfer and use, encompassing a much wider scope than only privacy issues.

Similarly to the GDPR, Brazil has recently approved the General Data Protection Law (abbreviated in Portuguese as LGPD), which will come into effect in early 2020. It is a comprehensive regulation, contemplating also extraterritorial activities involving Brazilian citizens [13]. Its concept of personal data refers to "information related to an identified or identifiable person". It is similar to the GDPR definition, with wording susceptible to a broad spectrum of interpretations. A more important issue, though, is its relation – complementarity and conflict – with the Access to Information Law that came into effect in 2012, and since then has fostered the publication of an extraordinary volume of government data.

Privacy frameworks have been proposed independently of the GDPR. The National Institute of Standards and Technology (NIST), in collaboration with partners from the private and public sectors, is developing a privacy framework, intended to be compatible with US and international regulations. It includes a catalog of issues, focusing on privacy risk management. Another example is the Privacy Control Framework (PCF) of NOREA (Dutch Association of chartered IT-auditors) [20], proposed as a privacy control assessment guide. It contains 104 controls in total, divided over 32 subjects in 9 Lifecycle Management phases. The Asia-Pacific Economic Cooperation (APEC) Cross-Border Privacy Rules (CBPR) [4] is a certification program designed to ensure the flow of personal information across APEC member borders. These frameworks, although comprehensive, are not intended to propose methods to deal with each privacy issue identified.
As a final observation, we note that privacy is concerned with controlling the use of information related to individual persons, whereas transparency is concerned with controlling the use of a broader set of information, including – but not limited to – information about persons.

3 A Summary of the OurPrivacy Framework

We present here a summary of the OurPrivacy framework, in order to describe the main concepts needed to characterize later the issues to be discussed.
OurPrivacy assumes that the information to be used is stored in a Knowledge Graph (KG), a term introduced by Google in 2012 to denote a graph of inter-related items. Stated another way, a Knowledge Graph represents a collection of interlinked descriptions of entities – real-world objects, events, situations or abstract concepts – where:

- Descriptions have a formal structure that allows both people and computers to process them in an efficient and unambiguous manner;
- Entity descriptions contribute to one another, forming a network where each entity represents part of the description of the entities related to it.[1]

Thus, we define Privacy as "controlled access to information related to some agent", where all the information is stored in a given KG.
Figure 1 shows a diagram of the use of information within a KG, represented by a Request for an Action over an Artifact made by some Actor (henceforth named an Agent). We assume Artifacts contain a set of statements, thus forming a subgraph of the KG. Since Privacy refers to actions over some information, an Authorization must be granted for this Request, according to the Rules set forth by stakeholders. Stakeholders include persons *related to* the artifact, and institutional agents such as "the State" (whose rules are stated as laws). Rules may be based (drawn) on any information available in the KG.
For any Action request there may be several applicable rules, whose evaluation outcomes may result in conflicting Authorization responses. Since the Action request must have a definitive Authorization value, a conflict resolution strategy must be employed to reach a final decision. Such strategy is in turn subject to Governance Rules.
The general conflict resolution process can be expressed by the following algorithm.

1. Given Request(Agent, Action, Artifact),
   Let RS <- RuleSet(Artifact).
2. Let RS <- Sort-by-Precedence(RS, decreasing).
3. Let A <- DefaultAuthorization.
4. For each R in RS,
   a. Let AR <- Eval(R, Person, Action, Artifact);
   b. If AR = "Allowed" or AR = "Denied",
      return AR.
5. return A.

We first give an overall explanation of the algorithm, followed by a thorough discussion of some important details. In line 1, `RuleSet(Artifact)` is a function that computes the set of applicable rules. In line 2, this

[1] https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

`RuleSet` is sorted in a decreasing precedence order, according to some criteria. Line 3 establishes the `DefaultAuthorization` – i.e., "everything is allowed unless denied," or vice-versa. Next, each Rule is evaluated in the sorted order. The evaluation of a rule may result in "`Allowed`", "`Denied`" or "`Nil`" (i.e., undefined). Since the rules are ordered, the first one to obtain a non-Nil value is the Authorization result.
This algorithm abstracts the essential decisions that must be made, in lines 1-3, to wit:

1. Who can formulate a rule for a given artifact?
2. How are conflicts between rules resolved?
3. What are the allowed actions over artifacts?

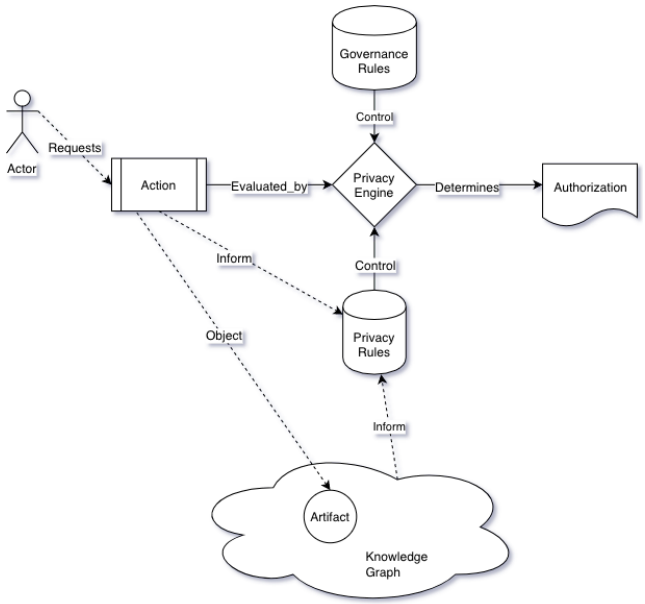


Figure 1: The Privacy Process

We detail possible answers to these questions in the following sub-sections.

3.1 Artifact Representation

We propose to represent an Artifact as a nanopublication[2]. "A nanopublication is the smallest unit of publishable information: an assertion graph about anything that can be uniquely identified and attributed to its author. Individual nanopublications can be cited by others and monitored for their impact on the Community". It is composed of three elements (a.k.a. named graphs): (i) an *assertion graph*, representing a minimal unit of thought that expresses a relationship (i.e., a predicate) between a subject and an object; (ii) a *provenance graph*, representing contextual metadata about the assertion, denoting "how the assertion came to be", including for example methods that were used to generate the assertion, authors, institutions, time-stamps, grants, links to DOIs, and URLs about the assertion; and (iii) a *publication info graph*, representing metadata about the nanopublication as a whole. The publication info differs from the provenance graph since the former pertains to both the assertion and provenance, while the latter relates just to the assertion.
As any nanopublication, an Artifact comprises an *assertion graph*, a *provenance graph* and a *publication info* graph, as follows.
The *assertion* graph will contain a set of assertions about the content of the Artifact, thus enabling the semantic evaluation of the privacy rules. The assertions in this graph are a subset of the assertions in the KG.

[2] http://nanopub.org

Thus, if an Agent posts an image that depicts some event, it would have statements about the participants and their roles, location information, date information, etc. as assertions in the assertion graph. We do not discuss here how such assertions containing semantic information are obtained (extracted) from the representation (image, video, text).
The *provenance* graph of the Artifact will contain provenance information about the assertions in the assertion graph (e.g., what image or natural language processing software was used, recorded location info, whether the assertions were inferred using some inference engine, etc.). The provenance graph can be used to represent, to the desired level of detail, the supporting information for the assertions. For example, if an automated face recognition algorithm was used, the provenance information represented in the *provenance* graph of the nanopublication may inform which algorithm, which parameters were used in this particular case, and a confidence factor. In the case of a rule stated by a person that has legal authority over another (e.g., a parent over a child), the provenance information can include supporting evidence to establish such legal authority (e.g., a reference to a birth certificate that states that rule author is indeed the parent of child depicted in the photo).
The *publication info* graph will contain metadata about the creation of the Artifact itself, such as its creator.

3.2 Rules

The first decision is to determine what the applicable rules are given an Artifact, defined in the `RuleSet(Artifact)` function. This leads to the question "who has the right to define a Privacy Rule that controls actions over this artifact?" The definition itself indicates that it must be any agent that is somehow *related* to the information contained in the artifact; but, stated in this way, it is too general. Any useful instantiation of the framework must spell out such relation types, which can include:

- Any agent directly identified in the artifact – for example, some person appearing in a posted photo or video;
- Any agent referred to (mentioned) in the artifact – for example, some person cited in a post;
- The creator of the artifact;
- Any agent related to the creation of the artifact – for example the author of a video posted by someone else;
- Any agent who has legal jurisdiction over an agent identified or mentioned in the artifact;
- The legal system(s) that has(have) jurisdiction over the KG, the agent or the action request.

The presence of such relations can directly occur in the KG (i.e., as a typed edge), or as a composition of valid relations. Furthermore, the rules may (or may not) allow the use of inferred relations in the KG.
Rules are of the form antecedent => consequent, both of which are sets of statements [26]. The antecedent of privacy rules may refer to any statements in the KG, including

- Any statements in the graphs in the artifact's nanopublication (including provenance information);
- The identity of the agent requesting permission;
- The type of action;
- Information in the KG serving as contextual information), such as
    - Date/time of the request;
    - Current location of agents involved;

Rules are themselves Artifacts, so they are nanopublications. The actual specification of the rule is given in its assertion graph, using a notation such as N3Logic [26] or SWRL[3]. Governance Rules are Rules that include other Rules (in either antecedent or consequent), so in this sense they are meta-rules.
Governance rules can be used to express precedence relations between rules, and to capture certain aspects of the privacy process. A metarule can, among other things, control the creation of any rule. This can be used

[3] https://www.w3.org/Submission/SWRL/

to precisely control, for example, what are the relation types allowed to be considered when computing the generic concept of "information *related* to a person" in the definition of Privacy. The example below illustrates how one can restrict privacy rules to refer only to PersonalInformation (assuming the default authorization value is "Denied").

KG

PrivacyRule subClassOf Rule.

{?KG log:semantics ?KGS;
?KGS log:includes {?p1 ?r ?p2.
(?p1 rdf:type Person OR ?p2 rdf:type Person)}} => {?r rdf:type PersonalInformationRelation}

Meta Rule1

{?r1 type PrivacyRule, antecedent ?Ant. ?Ant log:semantics ?AS. ?AS log:includes {?p1 ?r ?p2}. ?r rdf:type PersonalInformationRelation}, ?act rdf:type Create, object:?r1}
=>{<at> type:authorization, rule <MRule1>, action <?act> , value "Allowed"}

The OurPrivacy Framework can be applied for transparency rules as well by simply allowing rules to refer to Artifacts of any kind, and not only Persons. This will be illustrated in the examples in Section 4.

The provenance information and the publication info associated to a metarule nanopubulication can provide support for accountability. The provenance sub-graph can include reference(s) to the pertinent legislation, as well as legal cases that support the interpretation encoded in the rule, and the publication info subgraph contains information about the rule author.

3.3 Conflict resolution

In the conflict resolution algorithm, the enabled rules are sorted in descending precedence order in the `Sort-by-Precedence(RS, decreasing)` function call. Given that, in general, many applicable rules regarding the authorization for the intended action may exist, it is possible that two different rules give mutually exclusive authorizations. This is resolved using a sorting function that typically combine several types of information to establish order relations between rules.

Several possible complementary order relations that can be employed, such as:

- Establish a hierarchy between users – rules defined by a higher-ranked user take precedence over rules defined by lower-ranked ones. For example, rules established by laws take precedence over rules stated by individuals;
- Establish a hierarchy over the relation types. Rules defined by users related to the artifact through a higher-ranked relation type take precedence over rules defined by users related to the artifact via a lower-ranked relation type. For example, one may state that "identification" takes precedence over "mentioning". Thus, in a video where a person A is identified, and person B is mentioned in a conversation, person A's rules would take precedence over person B's rules.
- Since hierarchies are partial orders, they may not completely define precedence, so further conflict resolution strategies are still needed. Such and Criado [11] identified six categories of strategies that can be employed. Most require user involvement at runtime, but the aggregation-based class can be easily incorporated into an algorithm. Strategies in this class define an aggregation function such as consensus, majority, minimum fixed number of votes, permit-overrides, deny-overrides, etc. – see also [9] – and replace the set of conflicting rules by a single aggregated result;

An alternative to the aggregation approach is to decompose the artifact into finer-grained elements so that each is subject to only one rule. This makes sense when the Action to be performed can be performed on each element independently, such as blurring the face of a given person in a group photo.

4 Example Scenarios and discussions

4.1 Transparency x Privacy

The first scenario regards the Brazilian Transparency Law, which states that information of public interest should be openly released. In this example, P1 is a congressperson and P2, a non-public person who financially contributed to P1's campaign. P1 formulated a rule (Rule 1) that would protect his contributions by allowing only government organizations to have access to them, i.e., the general public would not have permission to access them.

Nonetheless, because P1 is a congressperson, the Law of Transparency demands her/his campaign information to be publicly available. Given that laws have precedence over personal rules, the Rule representing the Law of Transparency (Rule 2) precedes Rule 1. The following statements and rules characterize this scenario, where we assume the default authorization is "`Denied`".

KG

Read subClassOf Action.

<t> type Transaction, recipient <P1>.

<t> type PublicInterestInformation.

Rule 1.

{?T type Transaction; assertions ?Assrt. ?Assrt log:semantics ?AS. ?AS log:includes {?T category <CampaignContribution>. <act> type Read; object ?T. ?Org intends <act>, org:subOrgOf <Government>. ?T author <P1>}

=> {<at> type Authorization, rule "Rule1", action <act>, value "Allowed"}

Rule 2.

{?pi type PublicInterestInformation. <act> type Read; object ?pi. ?Agent intends <act>}

=> {<at> type Authorization, rule "Rule2", action <act>, value "Allowed"}

The Transaction object `?T` representing the campaign contribution has the type `PublicInterestInformation` in the Knowledge Graph. According to Rule 2, which precedes Rule 1, everyone who intends to read such information should be allowed to do so.

A slightly different situation arises when P1 is actually the brother of P2, for example. This family relationship can be protected by the Brazilian Data Protection Law, which covers personal data (Rule 3). The `siblingOf` relationship in this case is personal information, but it can also be regarded as public interest information. This entails a discussion on whether public interest information precedes personal data protected by the Brazilian regulation, as it is unclear which law should be applied.

The fact that P1 is sibling of P2 may not be public knowledge, or present in the KG, but the fact that P2 contributed to P1's campaign is. So, does the fact that they are siblings matter in this case? The General Data Protection Regulation states an individual can restrict actions on his/her personal information. If there is a public interest information involved, such as the financial contribution, then should other personal information become that of public interest, and thus subject to the transparency legislation?

In the statements below, we consider the sibling relationship as being both personal information and public interest information, which would then generate direct conflict based on Rule 2, stated above, and Rule 3, below.

KG

Read subClassOf Action.

<t> type Transaction, recipient <P1>.

<t> type PublicInterestInformation.

<P1> siblingOf <P2>

<siblingOf> type rdf:Property, PersonalInformation, PublicInterestInformaton.

Rule 3.
{?pi type PersonalInformation. <act> type Read; object ?pi. ?Agent intends <act>}
=> {<at> type Authorization, rule "Rule3", action <act>, value "Denied"}

Taken separately, Rule 2 and Rule 3 may act as default rules, in terms of the conflict resolution algorithm. Default rules are superseded when a more specific rule precedes the default rule. However, in this example these two rules are in conflict with one another. The precedence between them is arguable, because, as discussed, it is unclear whether personal information precedes public interest information or vice-versa. Hence, whether the `siblingOf` relationship between P1 and P2 could be acted upon or not depends entirely on the implementation of the `Sort-By-Precedence` function.

4.2 Ensuring transparency for public citizens: The right to be forgotten case

This example takes place on Twitter and features a famous person or politician, such as the President of Brazil – Jair Bolsonaro – as the main participant. Imagine a citizen called Lucas, who automatically captures Bolsonaro's tweets and re-posts them on his Twitter timeline. If Bolsonaro deletes a tweet he had previously posted, Lucas will still have the original tweet published on his timeline.

This raises an interesting question regarding the right to be forgotten. Would it be possible for someone (in this case, Bolsonaro) to specify a rule that would prevent others (in this case, Lucas) from re-posting his tweets? Is it legitimate that a user prevents others from sharing his/her own words, or pictures, especially if they have been deleted? Also, in a social network setting, it is virtually impossible to track when users copy and paste a text, or download and upload an image, without referring to the original content.

The possible solutions may also be deemed controversial. It is still not clear whether a public citizen such as Bolsonaro should have the ability to prevent some other user from publishing his original content. Moreover, the right to be forgotten is not guaranteed when it is possible for users to manually re-post tweets or posts.

KG
Read subClassOf Action.
<tw1> type Tweet, author <JairBolsonaro>.
<tw2> type Tweet, author <Lucas>, copyOf <tw1>.
<JairBolsonaro> type PublicFigure.

Rule 1.
{?T type Tweet; <act> type Read, object ?T. ?Person intends <act>.
?T author <JairBolsonaro>}
=> {<at> type Authorization, rule "Rule1", action <act>, value "Denied"}

Rule 2.
{?T1 type Tweet; <act> type Read, object ?T1; ?Person intends <act>.
?T type Tweet; ?T copyOf ?T1. ?T1 author <JairBolsonaro>}
=> {<at> type Authorization, rule "Rule2", action <act>, value "Denied"}

Rule 3.
{<act> type Read. ?Person intends <act>.
?T type Tweet, author <Lucas>}
=> {<at> type Authorization, rule "Rule3", action <act>, value "Allowed"}

Rule 1's antecedent includes the `author` property, indicating that this property is understood as a valid "related to" instance with respect to the definition of privacy. Although not shown, this would be stated in a meta-rule that allows creating rules with this property in its antecedent. By this rule, Jair Bolsonaro has denied Read actions on any of his tweets. We assume Jair Bolsonaro has established this rule at a moment in time after Lucas has copied his tweets, producing the same effect as deleting the tweets (at least for the purposes of the example).

Similarly, by Rule 2, the use of the composition of `author` and `CopyOf` properties in its antecedent expresses the understanding that a copy of someone's tweet is "related to" its original author, and therefore

<JairBolsonaro> is entitled to establish it as a privacy rule. Thus Rule 3 denies Read actions on any copy of his tweets.

Rule 3 (established by Lucas) simply states that all his tweets can be read by everybody. If the default Authorization in this scenario is "Allowed", this rule would be redundant.

In addition to these rules, assume the KG has the statements shown under "KG". If a person intends to read tw2, Rule 2 will return a "Denied" authorization, whereas Rule 3 will return an "Allowed" authorization.

If it is interpreted that a public figure is not protected under the Right to be Forgotten law, since Jair Bolsonaro is a public figure, then Rules 1 and 2 would not be allowed to be created by the Governance Rules, and Rule 3 would prevail. Conversely, if this tweet is, for example, about an offense that he made to someone in the past, it could be considered a proof of a crime. In this case, another rule (based on the Law) would prevail. So, do we (public persons or not) really have the right to be forgotten?

In a scenario where there are no such laws, other criteria may be reflected in the conflict resolution rules. For instance, based on an interpretation of the privacy legislation, Rule 2 would precede Rule 3, and the read action would be denied.

4.3 Privacy issues on digital memories: The Black Mirror case

Futuristic television series *Black Mirror*[4] presents interesting scenarios of privacy invasion. In "The entire history of you", the final episode of the first season, which first aired on December 2011, people use a device called Grain, which is a memory implant that records everything their eyes see, making it possible to browse through previous memories as if they were videos. The main character of the episode, Liam, threatens other people's security to gain access to old memories stored on their Grain [6]. Liam pressures his wife, Ffion, into showing him a memory she has of a sexual affair she had during their marriage. In the episode, she tries to delete the memory, but he notices it and violently prevents the deletion.

In this scenario, Ffion has no protection over her memories. Although she is the one who actively controls the system, if threatened, there is no mechanism that blocks or prevents unwanted access. With OurPrivacy, though, her privacy could be sustained with the help of rules.

For this example, it would make sense for Ffion to write a rule preventing her husband from accessing memories of her love affair (Rule 1). We can consider that each piece of memory is a nanopublication. Since all memories are recorded by the same device, at least part of the provenance information should be the same. The date, time and location, however, will change accordingly.

In Ffion's rule for this example, the assertion graph is key. She is concerned about memories of her affair, which means that memories containing romantic or sexual events with her lover should not be accessed by Liam, her husband. The content of each memory is stated in its assertion graph. Because of that, when Liam states that he wants to see a specific memory – which, in our framework, is an action –, the system checks in the assertion graph of that memory if it contains any mention of a sexual or romantic event with Ffion's lover. If it does, because of Ffion's rule, Liam is denied access to that memory.

Rule 1.

{?M type Memory; assertions ?Assrt. ?Assrt log:semantics ?AS. ?AS log:includes {<Jonas> sexualActWith <Ffion>}. <act> type Read, object ?M. <Liam> intends <act>.

?M author <Ffion>}

=> {<at> type Authorization, rule "Rule1", action <act>, value "Denied"}

Because Ffion is the author of the memories in question, she has the capability of writing rules about them. However, since she is married to Liam and memories of infidelity can directly affect their marriage, it can be debated whether Liam should have the ability to write rules concerning these memories as well. This would cause a conflict, as the values of the ultimate authorization would differ. Liam's rule of wanting access to Ffion's adultery would directly clash with Ffion's rule that prevents Liam's access.

[4] https://www.netflix.com/title/70264888

If the couple had any sort of prenuptial agreement or other type of contract that contained an article regarding infidelity, this document could be used as a conflict resolution artifact. Legal papers can play a great part in our framework, acting as a means to solve conflicts according to the implementation of the `Sort-by-Precedence` function, as discussed in Section 2.

If the prenuptial agreement in question contained a specific stipulation of what would happen to the couple's memories in case of adultery, this could be translated into Rule 2. The clause is that, in case of Ffion having sexual relations with a person other than Liam, and Liam wanting to see that memory, Liam would be allowed to do so.

KG

Read subClassOf Action.

<m1> type Memory, author <Ffion>.

<prenup> type LegalDocument, author <Lawyer1>

Rule 2.

{?M type Memory; assertions ?Assrt. ?Assrt log:semantics ?AS. ?AS log:includes {?Person sexualActWith <Ffion>}. ?Person differentFrom <Liam>. <act> type Read, object ?M. <Liam> intends <act>. ?M author <Ffion>. "Rule2" provenance ?Prov; ?Prov log:semantics ?PV. ?PV log:includes {?prenup type prov:Entity, LegalDocument}}

=> {<at> type Authorization, rule "Rule2", action <act>, value "Allowed"}

Rule 1 and 2 have a clear conflict. Given the legal framework, and the fact that Rule 2's provenance is a legal document, Rule 2 precedes Rule 1, and thus the conflict resolution algorithm will return "`Allowed`" for the authorization request.

5 Conclusions

This paper illustrates the use of OurPrivacy as a framework of reference to address two issues that are commonly taken as contradictory: that of public transparency versus individual privacy We show that, essentially, they refer to opposite sides of the same Quality spectrum and, therefore, may be accounted as two faces of the same coin.

We use the OurPrivacy Framework as a conceptual framework of reference to precisely characterize where in the design process the decisions must be made to handle both transparency and privacy concerns, and how it can support accountability. We show and discuss several real-life scenarios, which evidence this conclusion.

As counter-intuitive as it may seem at first, this is a very interesting conclusion, which may impact the definition of common mechanisms to ensure compliance of both issues without leading to contradictory and non-viable scenarios.

We emphasize that, in its current stage as a conceptual framework, this work does not address concerns with implementation. One of the main goals is to provide support on how to assess the trade-off between privacy and transparency through a sort-by-precedence rule. In this sense, the framework can use any strategy already proposed in the literature, as so, we do not claim to propose a new strategy of conflict resolution.

As the interplay between rules can become quite complex, as future work, we envisage developing a simulator based on OurPrivacy to help designers anticipate the impact of their decisions using the framework, in a way similar to [13].

The issues and concerns raised with OurPrivacy must be dealt with in any implementation. This poses significant challenges that we are already addressing as part of ongoing and future work.

ACKNOWLEDGMENTS

Daniel Schwabe and Flávia Santoro were partially supported by a grant from CNPq.

REFERENCES

[1] Adam D. Moore. 2003. Privacy: Its Meaning and Value. *American Philosophical Quarterly* 40, 3 (2003), 215–227.

[2] Alan F Westin. Privacy and Freedom. Washington and Lee Law Review 25, 1, 16.

[3] Albert Meijer. 2013. Understanding the Complex Dynamics of Transparency. Public Administration Review 73, 3 (2013), 429–439. DOI:https://doi.org/10.1111/puar.12032
[4] APEC Cross Border Privacy Enforcement Arrangement. Retrieved February 27th 2019 from https://cbprs.blob.core.windows.net/files/Cross%20Border%20Privacy%20Enforcement%20Arrangement.pdf
[5] Bianca Rodrigues Teixeira, Daniel Schwabe, Fernanda Baião, Flavia Maria Santoro, Carlos Laufer, Sérgio Lifschitz, Rosa M. Costa2 Simone D. J. Barbosa. 2019. OurPrivacy: A Framework for Privacy in Social Media, submitted for publication, WWW Journal.
[6] Bianca Rodrigues Teixeira and Flávia Maria Santoro. 2017. Memory and Privacy in The Entire History of You. In Proceedings of Workshop Re-coding Black Mirror 2017 Workshop - 16th International Semantic Web Conference (ISWC 2017).
[7] Daniel J. Solove. 2010. Understanding Privacy (2/28/10 edition ed.). Harvard University Press, Cambridge, MA.
[8] Ethan S. Bernstein. 2016. Making Transparency Transparent: The Evolution of Observation in Management Theory. ANNALS 11, 1 (October 2016), 217–266. DOI:https://doi.org/10.5465/annals.2014.0076
[9] Federica Paci, Anna Squicciarini, and Nicola Zannone. 2018. Survey on Access Control for Community-Centered Collaborative Systems. ACM Computing. Surveys. 51, 1 (January 2018), 6:1–6:38. DOI:https://doi.org/10.1145/3146025
[10] George Ritzer. 2008. Transparency in Global Change: The Vanguard of the Open Society by Burkart Holzner and Leslie Holzner. American Journal of Sociology 114, 1 (July 2008), 267–269. DOI:https://doi.org/10.1086/592526
[11] Jose M. Such and Natalia Criado. 2018. Multiparty Privacy in Social Media. Commun. ACM 61, 8 (July 2018), 74–81. DOI:https://doi.org/10.1145/3208039
[12] Klaus Schwab. 2017. The Fourth Industrial Revolution. Currency, New York.
[13] Manoel Pereira Junior, Simone Isabela de Rezende Xavier, and Raquel Oliveira Prates. 2014. Investigating the Use of a Simulator to Support Users in Anticipating Impact of Privacy Settings in Facebook. In Proceedings of the 18th International Conference on Supporting Group Work (GROUP '14), 63–72. DOI:https://doi.org/10.1145/2660398.2660419
[14] Manuel Castells. 2000. The Rise of the Network Society (2nd ed.). Blackwell Publishers, Inc., Cambridge, MA, USA.
[15] Mark Bovens 2007, Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal 13, 4 (2007) 447-468.
[16] Mathias Klang and Andrew Murray. 2005. Human Rights in the Digital Age. Psychology Press.
[17] Maximilian Heimstädt and Leonhard Dobusch. 2018. Politics of Disclosure: Organizational Transparency as Multiactor Negotiation. Public Administration Review 78, 5 (2018), 727–738. DOI:https://doi.org/10.1111/puar.12895
[18] Michael Johnston. 2014. Transparency. Encyclopedia Britannica. Retrieved February 4, 2019 from https://www.britannica.com/topic/transparency-government
[19] Michael Schudson. 2015. Politics and the Culture of Transparency. In The Rise of the Right to Know. Harvard University Press, Cambridge, MA, 1945–1975.
[20] NOREA Guide Privacy Control Framework - Control objectives and controls for privacy audits and privacy assurance engagements. Retrieved February 27th, 2019 from https://www.norea.nl/download/?id=4160
[21] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[22] Renato Monteiro. 2018. The new Brazilian General Data Protection Law — a detailed analysis. International Association of Privacy Professionals. Retrieved February 4, 2019 from https://iapp.org/news/a/the-new-brazilian-general-data-protection-law-a-detailed-analysis/
[23] Serge Gutwirth, and Paul De Hert 2007. Privacy, Data Protection and Law Enforcement - Opacity of the Individual and Transparency of Power. Privacy and the Criminal Law, Erik Claes, Antony Duff, and Serge Gutwirth. eds., Antwerpen-Oxford: Intersentia, 61-104.

[24] Sharon S. Dawes and Natalie Helbig. 2010. Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency. In Electronic Government (Lecture Notes in Computer Science), 50–60.
[25] Thomas N. Hale. 2008. Transparency, Accountability, and Global Governance. Global Governance 14, 1 (2008), 73–94.
[26] Tim Berners-lee, Dan Connolly, Lalana Kagal, Yosi Scharf, and Jim Hendler. 2008. N3Logic: A Logical Framework for the World Wide Web. Theory Pract. Log. Program. 8, 3 (May 2008), 249–269. DOI:https://doi.org/10.1017/S1471068407003213
[27] United Nations, Universal Bill of Human Rights, Resolution A/RES/217(III)[A]. Retrieved February 4, 2019 from http://unbisnet.un.org:8080/ipac20/ipac.jsp?session=14O243550E15G.60956&profile=voting&uri=full=3100023~!909326~!67
[28] W. A. Parent. 1983. Privacy, Morality, and the Law. Philosophy & Public Affairs 12, 4 (1983), 269–288.
[29] William H. Dutton. 2009. The Fifth Estate Emerging Through the Network of Networks. Prometheus 27, 1 (2009), 1–15.
[30] Yvonne McDermott. 2017. Conceptualising the right to data protection in an era of Big Data. Big Data & Society 4, 1 (June 2017), 2053951716686994. DOI:https://doi.org/10.1177/2053951716686994

Achieving Differential Privacy and Fairness in Logistic Regression

Depeng Xu
University of Arkansas
depengxu@uark.edu

Shuhan Yuan
University of Arkansas
sy005@uark.edu

Xintao Wu
University of Arkansas
xintaowu@uark.edu

ABSTRACT

Machine learning algorithms are used to make decisions in various applications. These algorithms rely on large amounts of sensitive individual information to work properly. Hence, there are sociological concerns about machine learning algorithms on matters like privacy and fairness. Currently, many studies focus on only protecting individual privacy or ensuring fairness of algorithms. However, how to meet both privacy and fairness requirements simultaneously in machine learning algorithms is under exploited. In this paper, we focus on one classic machine learning model, logistic regression, and develop differentially private and fair logistic regression models by combining functional mechanism and decision boundary fairness in a joint form. Theoretical analysis and empirical evaluations demonstrate our approaches effectively achieve both differential privacy and fairness while preserving good utility.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → **Law, social and behavioral sciences**;

KEYWORDS

Differential Privacy; Fairness-aware Learning; Logistic Regression

ACM Reference Format:
Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving Differential Privacy and Fairness in Logistic Regression. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317584

1 INTRODUCTION

Nowadays, machine learning algorithms are being widely used to automatically make decisions, such as loan application and student admission, based on our individual information. It is important to address individuals' sociological concerns such as privacy and fairness and meet government laws and regulations (e.g., General Data Protection and Regulation on data protection and privacy, and Fair Credit Reporting Act or Equal Credit Opportunity Act on fairness) in training and deploying machine learning algorithms [2, 12].

Differential privacy has been established as a standard privacy model to achieve opt-out right of individuals [6]. Generally speaking, differential privacy guarantees the query results or the released model cannot be exploited by attackers to derive whether one particular record is present or absent in the underlining dataset. Various mechanisms have been proposed to achieve differential privacy [4]. For example, the Laplace mechanism works by injecting random noise into the released results such that the inclusion or exclusion of a single individual record from the dataset makes no statistical difference to the results found [6]. For prediction models, objective perturbation [1] and functional mechanism [16], which add noise to the objective function rather than parameters of built models, have been shown great success.

Meanwhile, fairness-aware learning is increasingly receiving attention in the machine learning field. Many studies have shown that classification models have biased performance against the *protected group*, since the classifiers are only trained to maximize the prediction accuracy. Current research to achieve fair classification can be mainly categorized into two groups: in-processing methods which incorporate fairness constraints into the classification models [11, 15], and pre/post-processing methods which modify the training data and/or the potentially unfair predictions made by the classifiers [5, 8, 10, 17, 18].

In this work, we focus on how to achieve both differential privacy and fairness in logistic regression – a widely-used classification model. It's challenging to achieve both requirements efficiently. The goal of differential privacy in a classification model is to make sure the classifier output is indistinguishable whether an individual record exists in the dataset or not. Its focus is on the individual level. The goal of fairness-aware learning is to make sure that predictions of the protected group are identical to those of the unprotected group, e.g., admission rate of female (as protected group) should be same to male (as unprotected group). Its focus is on the group level. There are few studies on achieving both privacy protection and fairness. Research in [5] proposed a notion of fairness that is a generalization of differential privacy. Research in [9] developed a pattern sanitization method that achieves k-anonymity and fairness. Most recently, the position paper [7] argued for integrating recent research on fairness and non-discrimination to socio-technical systems that provide privacy protection. However, there is no formal study on how to achieve both differential privacy and fairness in classification models.

We develop two methods to achieve differential privacy and fairness in logistic regression. Our simple method incorporates the decision boundary fairness constraint into the objective function of the logistic regression as a penalty term and then applies the functional mechanism to the whole constrained objective function to achieve differential privacy. The decision boundary fairness

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317584

constraint of logistic regression is defined as the covariance between the users' protected attribute and the signed distance from the users' unprotected attribute vectors to the decision boundary, and can be further formulated as the signed distance between the centroids of the protected and unprotected groups. To achieve differential privacy, the functional mechanism brings randomness to the polynomial coefficients of the constrained objective function by introducing Laplace noise with zero mean. Because the penalty term contributes to the global sensitivity of objective function, this simple approach may inject too much noise to the objective function, which reduces the utility of the built logistic regression model. We further develop an enhanced model that injects Laplace noise with shifted mean to the objective function of logistic regression. Our idea is based on the connection between ways of achieving differential privacy and fairness. We notice that both the fairness constraint and functional mechanism perturb the polynomial coefficients of the original objective function. Hence, we can combine them as a single term. In fact, the decision boundary fairness constraint of logistic regression can be treated as a shift of the polynomial coefficients by the signed distance between the centroids of the protected and unprotected groups. As a result, we add noise from a Laplace distribution with non-zero mean that is derived from the fairness constraint. In this way, the fairness constraint is not a penalty term, so we can use privacy budget more efficiently and add less noise.

Our contributions are as follows: 1) To our best knowledge, this is the first work to study how to achieve both differential privacy and fairness in classification models. 2) We develop two methods to achieve differential privacy and fairness in logistic regression. In particular, our enhanced method, which adds Laplace noise with non-zero mean as equivalence to fairness constraint, can reduce the amount of added noise and hence better preserve utility. 3) We conduct evaluation on two real-world datasets and results show that our approaches meet both differential privacy and fairness requirements while achieving good utility.

2 PRELIMINARY

Let $D = \{\mathbf{X}, S, Y\}$ be a dataset with n tuples $t_1, t_2, \cdots, t_n$, where $\mathbf{X} = (X_1, X_2, \cdots, X_d)$ indicates d unprotected attributes; S denotes the protected attribute; Y is the decision. For each tuple $t_i = \{\mathbf{x}_i, s_i, y_i\}$, without loss of generality, we assume $x_{i(l)} \in [0, 1]$ for $l = (1, 2, \cdots, d)$, $s_i \in \{0, 1\}$, and $y_i \in \{0, 1\}$. Our objective is to build a classification model $\hat{y} = q(\mathbf{x}; \mathbf{w})$ with parameter $\mathbf{w}$ from D that achieves reasonable utility and meets both fairness and differential privacy requirements. To fit $\mathbf{w}$ to make accurate predictions, we have an objective function $f_D(\mathbf{w}) = \sum_{i=1}^{n} f(t_i; \mathbf{w})$ that takes t_i and $\mathbf{w}$ as input. The optimal model parameter $\bar{\mathbf{w}}$ is defined as: $\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} f(t_i; \mathbf{w})$.

2.1 Differential Privacy

Differential privacy guarantees the output of a query q be insensitive to the presence or absence of any one individual record in a dataset. We use D and D' to denote two neighboring datasets which differ in exactly one record.

Definition 2.1. **Differential Privacy** [6]. A mechanism $\mathcal{M}$ satisfies ε-differential privacy, if for all neighboring datasets D and D' and all subsets Z of $\mathcal{M}$'s range:

$$\Pr(\mathcal{M}(D) \in Z) \leq \exp(\varepsilon) \cdot \Pr(\mathcal{M}(D') \in Z). \quad (1)$$

The parameter ε denotes the privacy budget (smaller values indicate a stronger privacy guarantee).

Definition 2.2. **Global Sensitivity** [6]. Given a query $q: D \to \mathbb{R}^d$, the global sensitivity Δ is defined as $\Delta = \max_{D, D'} ||q(D) - q(D')||_1$.

The global sensitivity measures the maximum possible change in $q(D)$ when one record in the dataset changes. The Laplace mechanism is a popular method to achieve differential privacy. It adds identical independent noise into each output value of $q(D)$.

Definition 2.3. **Laplace Mechanism** [6]. Given a dataset D and a query q, a mechanism $\mathcal{M}(D) = q(D) + \boldsymbol{\eta}$ satisfies ε-differential privacy, where $\boldsymbol{\eta}$ is a random vector drawn from $Lap(0, \frac{\Delta}{\varepsilon})$ [1].

Functional Mechanism. Functional mechanism [16] is a differentially private method designed for optimization based models. It achieves ε-differential privacy by injecting noise into the objective function of the model and returns privacy preserving parameter $\bar{\mathbf{w}}$ that minimizes the perturbed objective function.

Because the objective function $f_D(\mathbf{w})$ is a complicated function of $\mathbf{w}$, the functional mechanism exploits the polynomial representation of $f_D(\mathbf{w})$. The model parameter $\mathbf{w}$ is a vector that contains d values $w_1, w_2, \cdots, w_d$. Let $\phi(\mathbf{w})$ denote a product of $w_1, w_2, \cdots, w_d$, i.e., $\phi(\mathbf{w}) = w_1^{c_1} \cdot w_2^{c_2} \cdots w_d^{c_d}$ for some $c_1, c_2, \cdots, c_d \in \mathbb{N}$. Let $\boldsymbol{\Phi}_j$ ($j \in \mathbb{N}$) denote the set of all products of $w_1, w_2, \cdots, w_d$ with degree j, i.e., $\boldsymbol{\Phi}_j = \{w_1^{c_1} w_2^{c_2} \cdots w_d^{c_d} | \sum_{l=1}^{d} c_l = j\}$. For example, $\boldsymbol{\Phi}_1 = \{w_1, w_2, \cdots, w_d\}$, and $\boldsymbol{\Phi}_2 = \{w_i \cdot w_j | i, j \in [1, d]\}$.

Based on the Stone-Weierstrass Theorem [13], any continuous and differentiable function can be expressed in the polynomial representation. Hence, the objective function $f_D(\mathbf{w})$ can be expressed as a polynomial of $w_1, w_2, \cdots, w_d$, for some $J \in \mathbb{N}$:

$$f_D(\mathbf{w}) = \sum_{i=1}^{n} \sum_{j=0}^{J} \sum_{\phi \in \boldsymbol{\Phi}_j} \lambda_{\phi t_i} \phi(\mathbf{w}), \quad (2)$$

where $\lambda_{\phi t_i} \in \mathbb{R}$ denotes the coefficient of $\phi(\mathbf{w})$ in the polynomial.

Functional mechanism perturbs the objective function $f_D(\mathbf{w})$ by injecting Laplace noise into its polynomial coefficients $\bar{\lambda}_\phi = \sum_{i=1}^{n} \lambda_{\phi t_i} + Lap(0, \frac{\Delta}{\varepsilon})$, where $\Delta = 2 \max_t \sum_{j=1}^{J} \sum_{\phi \in \boldsymbol{\Phi}_j} ||\lambda_{\phi t}||_1$. Then the model parameter $\bar{\mathbf{w}}$ is derived to minimize the perturbed function $\bar{f}_D(\mathbf{w})$.

Applying Functional Mechanism on Logistic Regression. A logistic regression on D returns a function which predicts $\hat{y}_i = 1$ with probability:

$$\hat{y}_i = q(\mathbf{x}_i; \mathbf{w}) = \exp(\mathbf{x}_i^T \mathbf{w}) / (1 + \exp(\mathbf{x}_i^T \mathbf{w})). \quad (3)$$

The objective function of logistic regression is defined as:

$$f_D(\mathbf{w}) = \sum_{i=1}^{n} \left[\log(1 + \exp(\mathbf{x}_i^T \mathbf{w})) - y_i \mathbf{x}_i^T \mathbf{w} \right]. \quad (4)$$

As the polynomial form of $f_D(\mathbf{w})$ in Equation 4 contains terms with unbounded degrees, to apply the functional mechanism, Equation

[1] The Laplace distribution $Lap(\eta | \mu, b)$ with mean μ and scale b has probability density function $Lap(\eta | \mu, b) = \frac{1}{2b} \exp(\frac{|x - \mu|}{b})$. Its variance is $2b^2$.

4 is rewritten as the approximate polynomial representation based on Taylor expansion [16]:

$$f_D(\mathbf{w}) = \Big(\sum_{i=1}^{n} \sum_{j=0}^{2} \frac{f_1^{(j)}(0)}{j!} \left(\mathbf{x}_i^T \mathbf{w}\right)^j \Big) - \Big(\sum_{i=1}^{n} y_i \mathbf{x}_i^T \Big) \mathbf{w}, \quad (5)$$

where $f_1(\cdot) = \log(1 + \exp(\cdot))$.

When rewriting Equation 5 in the form of Equation 2, we have

$$\{\lambda_{\phi t_i}\}_{\phi \in \Phi_1} =: \lambda_{1t_i} = \Big(\tfrac{f_1^{(1)}(0)}{1!} \mathbf{x}_i \Big) - \Big(y_i \mathbf{x}_i \Big), \quad (6)$$

$$\{\lambda_{\phi t_i}\}_{\phi \in \Phi_2} =: \lambda_{2t_i} = \tfrac{f_1^{(2)}(0)}{2!} \left(\mathbf{x}_i\right)^2. \quad (7)$$

The global sensitivity of $f_D(\mathbf{w})$ is:

$$\begin{aligned} \Delta_f &= 2 \max_t \left(\left| \Big(\frac{f_1^{(1)}(0)}{1!} - y \Big) \sum_{l=1}^{d} x_{(l)} \right| + \left| \frac{f_1^{(2)}(0)}{2!} \sum_{l,m}^{d} x_{(l)} x_{(m)} \right| \right) \\ &\leq 2(\frac{d}{2} + \frac{d^2}{8}) = \frac{d^2}{4} + d. \end{aligned} \quad (8)$$

Thus, to achieve ε-differential privacy, the functional mechanism adds $Lap(0, \frac{\Delta_f}{\varepsilon})$ noise to the polynomial coefficients of the objective function.

2.2 Classification Fairness

There are many definitions or requirements of classification fairness. A common notion of group fairness is demographic parity, which requires that a decision Y is independent of the protected attribute S.

Definition 2.4. **Demographic Parity** [10]. A classification model $\hat{y} = q(\mathbf{x}; \mathbf{w})$ satisfies demographic parity if $\Pr(\hat{Y} = 1|S = 1) = \Pr(\hat{Y} = 1|S = 0)$.

The discrimination of the model can be quantified by risk difference (RD):

$$RD = |\Pr(\hat{Y} = 1|S = 1) - \Pr(\hat{Y} = 1|S = 0)|. \quad (9)$$

To achieve classification fairness, the in-processing approaches are to find parameter $\mathbf{w}$ that minimizes the objective function under a fairness constraint:

$$\begin{aligned} &\text{minimize} \quad f_D(\mathbf{w}) \\ &\text{subject to} \quad g_D(\mathbf{w}) \leq \tau,\ g_D(\mathbf{w}) \geq -\tau, \end{aligned} \quad (10)$$

where $g_D(\mathbf{w})$ is the constraint term; $\tau \in \mathbf{R}^+$ is the threshold of constraint. For example, in [15], the fairness constraint is defined as the covariance between the users' protected attribute and the signed distance from the users' unprotected attribute vectors to the decision boundary $\{d_{\mathbf{w}}(\mathbf{x}_i)\}_{i=1}^n$,

$$g_D(\mathbf{w}) = \mathbb{E}[(s - \bar{s}) d_{\mathbf{w}}(\mathbf{x})] - \mathbb{E}[(s - \bar{s})] d_{\mathbf{w}}(\mathbf{x}) \propto \sum_{i=1}^{n} (s_i - \bar{s}) d_{\mathbf{w}}(\mathbf{x}_i), \quad (11)$$

where $\bar{s}$ is the mean value of the protected attribute; $\mathbb{E}[(s - \bar{s})] = 0$. For linear classification models, like logistic regression or linear SVMs, the decision boundary is simply the hyperplane defined by $\mathbf{x}^T \mathbf{w} = 0$. Then, Equation 11 reduces to $g_D(\mathbf{w}) = \sum_{i=1}^{n} (s_i - \bar{s}) \mathbf{x}_i^T \mathbf{w}$. The decision boundary fairness is proven to be a notion of fairness that minimizes the surrogate risk difference [14].

3 DIFFERENTIALLY PRIVATE AND FAIR LOGISTIC REGRESSION

In this section, we first present a simple approach (PFLR) to achieve differentially private and fair logistic regression. Then we show it leads to an enhanced approach (PFLR*) that achieves the same level of differential privacy and fairness with more flexibility and less noise.

3.1 PFLR: A Simple Approach

One straightforward approach to achieve both differential privacy and fairness is to apply the functional mechanism to the objective function with fairness constraint $\tilde{f}_D(\mathbf{w})$. We consider the fairness constraint as a penalty term to the objective function. Then, the objective function ends up as:

$$\tilde{f}_D(\mathbf{w}) = f_D(\mathbf{w}) + \alpha |g_D(\mathbf{w}) - \tau|, \quad (12)$$

where α is a hyper-parameter to balance the trade-off between utility and fairness. We set $\alpha = 1, \tau = 0$ for ease of discussion. The theoretical analysis still holds if α and τ are set to other values.

For logistic regression, $f_D(\mathbf{w})$ is the objective function shown in Equation 4, and $g_D(\mathbf{w})$ indicates the decision boundary fairness constraint shown in Equation 11. We then rewrite $\tilde{f}_D(\mathbf{w})$ in the polynomial form based on Taylor expansion.

$$\tilde{f}_D(\mathbf{w}) = \Big(\sum_{i=1}^{n} \sum_{j=0}^{2} \frac{f_1^{(j)}(0)}{j!} \left(\mathbf{x}_i^T \mathbf{w}\right)^j \Big) - \Big(\sum_{i=1}^{n} y_i \mathbf{x}_i^T \Big) \mathbf{w} + \Big| \sum_{i=1}^{n} (s_i - \bar{s}) \mathbf{x}_i^T \mathbf{w} \Big|. \quad (13)$$

By transforming $\tilde{f}_D(\mathbf{w})$ in the form of Equation 2, we have $\tilde{\lambda}_{1t_i} = \lambda_{1t_i} + |(s_i - \bar{s})\mathbf{x}_i|$ and $\tilde{\lambda}_{2t_i} = \lambda_{2t_i}$, where λ_{1t_i} and λ_{2t_i} are defined in Equations 6 and 7, respectively.

The global sensitivity of $\tilde{f}_D(\mathbf{w})$ is:

$$\begin{aligned} \Delta_{\tilde{f}} &= 2 \max_t \left(\left| \Big(\frac{f_1^{(1)}(0)}{1!} - y + |s - \bar{s}| \Big) \sum_{l=1}^{d} x_{(l)} \right| + \left| \frac{f_1^{(2)}(0)}{2!} \sum_{l,m}^{d} x_{(l)} x_{(m)} \right| \right) \\ &\leq 2(\frac{3d}{2} + \frac{d^2}{8}) = \frac{d^2}{4} + 3d. \end{aligned} \quad (14)$$

The derived $\bar{\mathbf{w}}$ satisfies ε-differential privacy by applying Algorithm 1. Since the objective function contains the fairness constraint as a penalty term, the classification model also achieves fairness.

Algorithm 1 PFLR (Dataset D, objective function $f_D(\mathbf{w})$, fairness constraint $g_D(\mathbf{w})$, privacy budget ε)

1: Set $\tilde{f}_D(\mathbf{w})$ by Equation 12
2: Compute λ_{1t_i} and λ_{2t_i} by Equations 6 and 7
3: Set $\tilde{\lambda}_{1t_i} = \lambda_{1t_i} + |(s_i - \bar{s})\mathbf{x}_i|$ and $\tilde{\lambda}_{2t_i} = \lambda_{2t_i}$
4: Set Δ by Equation 14
5: Set $\bar{\lambda}_1 = \left(\sum_{i=1}^{n} \tilde{\lambda}_{1t_i} \right) + Lap(0, \frac{\Delta_{\tilde{f}}}{\varepsilon})$
6: Set $\bar{\lambda}_2 = \left(\sum_{i=1}^{n} \tilde{\lambda}_{2t_i} \right) + Lap(0, \frac{\Delta_{\tilde{f}}}{\varepsilon})$
7: Let $\bar{f}_D(\mathbf{w}) = \bar{\lambda}_1^T \mathbf{\Phi}_1 + \bar{\lambda}_2^T \mathbf{\Phi}_2$
8: Compute $\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \bar{f}_D(\mathbf{w})$
9: Return $\bar{\mathbf{w}}$

Theorem 3.1. *Algorithm 1 satisfies ε-differential privacy.*

Proof. Assume D and D' are two neighbouring datasets. Without loss of generality, D and D' differ in row t_r and t'_r. Δ is calculated by Equation 14. $\bar{f}(\mathbf{w})$ is the output of Line 7. We have

$$\begin{aligned}
\frac{\Pr\{\bar{f}(\mathbf{w})|D\}}{\Pr\{\bar{f}(\mathbf{w})|D'\}} &= \frac{\prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon\left\|\sum_{t_i\in D}\tilde{\lambda}_{\phi t_i}-\bar{\lambda}_\phi\right\|_1}{\Delta_{\tilde{f}}}\right)}{\prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon\left\|\sum_{t'_i\in D'}\tilde{\lambda}_{\phi t'_i}-\bar{\lambda}_\phi\right\|_1}{\Delta_{\tilde{f}}}\right)} \\
&\le \prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot\left\|\sum_{t_i\in D}\tilde{\lambda}_{\phi t_i}-\sum_{t'_i\in D'}\tilde{\lambda}_{\phi t'_i}\right\|_1\right) \\
&= \prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot\left\|\tilde{\lambda}_{\phi t_r}-\tilde{\lambda}_{\phi t'_r}\right\|_1\right) \\
&= \exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}\left\|\tilde{\lambda}_{\phi t_r}-\tilde{\lambda}_{\phi t'_r}\right\|_1\right) \\
&\le \exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot 2\max_t\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}||\tilde{\lambda}_{\phi t}||_1\right)=\exp(\varepsilon).
\end{aligned} \tag{15}$$

□

3.2 PFLR*: An Enhanced Approach

We further enhance the simple approach by incorporating the fairness constraint into the Laplace noise.

In Equation 13, the fairness constraint $g_D(\mathbf{w})=\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i^T\mathbf{w}$ can be considered as shifting the first degree polynomial coefficients of $\mathbf{\Phi}_1$ in the objective function by $\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$. Since $\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$ is the signed distance between the centroids of the protected and unprotected groups, the derived $\bar{\mathbf{w}}$ based on the shifted coefficients ensures that the centroids of the protected and unprotected groups have the same distance to the decision boundary. Thus, the decision boundary fairness is achieved.

Meanwhile, the functional mechanism adds Laplace noise to inject randomness to the polynomial coefficients of the objective function. Because $\Pr\{\bar{f}(\mathbf{w})|D\}$ depends on the probability of the noise distribution, the designed Laplace noise provides the property of differential privacy.

Following this observation, instead of applying the fairness constraint as a penalty term to the objective function, we design a new functional mechanism that incorporates fairness constraint into the Laplace noise. In particular, we shift the polynomial coefficients when adding Laplace noise. The shift is achieved by setting the mean of Laplace distribution. As $g_D(\mathbf{w})$ only affects $\mathbf{\Phi}_1$, we change the mean of Laplace distribution $\mu=\{\mu_{(l)}\}_{l=1}^{d}$ from 0 to $\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$ for the coefficients related with $\mathbf{\Phi}_1$, so it has the equivalent effect to the fairness constraint. Formally, we have $\mu_{(l)}=\sum_{i=1}^{n}(s_i-\bar{s})x_{(l)t_i}$. Because the fairness constraint is not a penalty term of the objective function, PFLR* has the same objective function as the regular logistic regression $f_D(\mathbf{w})$ (defined in Equation 5). The global sensitivity of PFLR* is $\Delta_f=d^2/4+d$ as shown in Equation 8.

Note that, given a dataset, $\mu=\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$ is fixed. As we also access data when calculating μ, a small part of privacy budget ε_g is used to calculate μ in a differentially private manner by Laplace mechanism (Algorithm 2 Line 2). The sensitivity of μ is

$$\Delta_g=2\max_t\left|\sum_{l=1}^{d}(s_{t_r}-\bar{s})x_{t_r(l)}\right|\le 2d. \tag{16}$$

We formalize our new functional mechanism with fairness constraint as Algorithm 2. We split the total privacy budget ε into two parts ε_f and ε_g. We first calculate the differentially private μ with the privacy budget ε_g (Lines 1-2). Then, we introduce Laplace noise $Lap(\mu,\frac{\Delta_f}{\varepsilon_f})$ to the polynomial coefficients of the objective function with the privacy budget ε_f (Lines 3-7). Note that we only add the shifted Laplace noise to coefficients with $\mathbf{\Phi}_1$. Finally, we derive the optimized $\bar{\mathbf{w}}$ according to $\bar{f}_D(\mathbf{w})$ (Line 8). Next we show PFLR* achieves ε-differential privacy.

Theorem 3.2. *Algorithm 2 satisfies ε-differential privacy.*

Proof. Assume D and D' are two neighbouring datasets. Without loss of generality, D and D' differ in row t_r and t'_r. Δ_f is calculated by Equation 8. $\bar{f}(\mathbf{w})$ is the output of Line 7. Adding Laplace noise with non-zero mean to coefficients still satisfies ε_f-differential privacy.

$$\begin{aligned}
&\frac{\Pr\{\bar{f}(\mathbf{w})|D\}}{\Pr\{\bar{f}(\mathbf{w})|D'\}} \\
&= \frac{\prod_{\phi\in\mathbf{\Phi}_1}\exp\left(\frac{\varepsilon_f\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\lambda_\phi-\mu\right\|_1}{\Delta_f}\right)\prod_{\phi\in\mathbf{\Phi}_2}\exp\left(\frac{\varepsilon_f\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\lambda_\phi\right\|_1}{\Delta_f}\right)}{\prod_{\phi\in\mathbf{\Phi}_1}\exp\left(\frac{\varepsilon_f\left\|\sum_{t'_i\in D'}\lambda_{\phi t'_i}-\lambda_\phi-\mu\right\|_1}{\Delta_f}\right)\prod_{\phi\in\mathbf{\Phi}_2}\exp\left(\frac{\varepsilon_f\left\|\sum_{t'_i\in D'}\lambda_{\phi t'_i}-\lambda_\phi\right\|_1}{\Delta_f}\right)} \\
&\le \prod_{\phi\in\mathbf{\Phi}_1}\exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\sum_{t'_i\in D'}\lambda_{\phi t'_i}\right\|_1\right) \\
&\quad\cdot\prod_{\phi\in\mathbf{\Phi}_2}\exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\sum_{t'_i\in D'}\lambda_{\phi t'_i}\right\|_1\right) \\
&= \prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\left\|\lambda_{\phi t_r}-\lambda_{\phi t'_r}\right\|_1\right) \\
&= \exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}\left\|\lambda_{\phi t_r}-\lambda_{\phi t'_r}\right\|_1\right) \\
&\le \exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot 2\max_t\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}||\lambda_{\phi t}||_1\right)=\exp(\varepsilon_f)
\end{aligned} \tag{17}$$

Using Laplace mechanism, Line 2 satisfies ε_g-differential privacy on calculating μ. Since $\varepsilon_f+\varepsilon_g=\varepsilon$, Algorithm 2 satisfies ε-differential privacy. □

Algorithm 2 PFLR* (Database D, objective function $f_D(\mathbf{w})$, fairness constraint $g_D(\mathbf{w})$, privacy budget $\varepsilon_f,\varepsilon_g$)

1: Set Δ_g by Equation 16
2: Calculate $\mu=\{\mu_{(l)}\}_{l=1}^{d}$ by $\mu_{(l)}=\sum_{i=1}^{n}(s_i-\bar{s})x_{(l)t_i}+Lap(0,\frac{\Delta_g}{\varepsilon_g})$
3: Compute λ_{1t_i} and λ_{2t_i} by Equation 6 and 7
4: Set Δ_f by Equation 8
5: Set $\bar{\lambda}_1=\left(\sum_{i=1}^{n}\lambda_{1t_i}\right)+Lap(\mu,\frac{\Delta_f}{\varepsilon})$
6: Set $\bar{\lambda}_2=\left(\sum_{i=1}^{n}\lambda_{2t_i}\right)+Lap(0,\frac{\Delta_f}{\varepsilon})$
7: Let $\bar{f}_D(\mathbf{w})=\bar{\lambda}_1^T\mathbf{\Phi}_1+\bar{\lambda}_2^T\mathbf{\Phi}_2$
8: Compute $\bar{\mathbf{w}}=\arg\min_{\mathbf{w}}\bar{f}_D(\mathbf{w})$
9: Return $\bar{\mathbf{w}}$

Comparison between PFLR and PFLR*. In PFLR, the fairness constraint term $g_D(\mathbf{w})$ contributes to the sensitivity of the polynomial coefficients of the objective function. In PFLR*, the fairness constraint is achieved by adding Laplace noise with non-zero mean value, so the sensitivity of the polynomial coefficient

is not related to $g_D(\mathbf{w})$. PFLR* uses separate budgets on objective function and fairness constraint, so it's more flexible to find good trade-offs among privacy, fairness and utility. The fairness constraint has a much smaller sensitivity than the objective function ($\Delta_g \ll \Delta_f$). Hence, we can allocate a relatively small privacy budget on calculating the fairness constraint with Laplace mechanism, the utility is still satisfactory. Then, more privacy budget can be used to the objective function, resulting in a smaller scale of noise. More concretely, we compare the amount of noise that is introduced to the two proposed approaches. The variance of $\lambda_{\phi t_i} \in \{\bar{\lambda}_1, \bar{\lambda}_2\}$ in PFLR is $2(\Delta_{\tilde{f}}/\varepsilon)^2$. Thus, the variance of total noise added in PFLR is $2(d^2+d)(\Delta_{\tilde{f}}/\varepsilon)^2$. On the other hand, the variance of $\lambda_{\phi t_i} \in \bar{\lambda}_2$ in PFLR* is $Var(\bar{\lambda}_2) = 2(\Delta_f/\varepsilon_f)^2$. Because PFLR* injects Laplace noise to both $\bar{\lambda}_1$ and μ, based on the law of total variance, the variance of $\lambda_{\phi t_i} \in \bar{\lambda}_1$ in PFLR* is $Var(\bar{\lambda}_1) = \mathbb{E}[Var(\bar{\lambda}_1|\mu)] + Var(\mathbb{E}[\bar{\lambda}_1|\mu]) = 2(\Delta_f/\varepsilon_f)^2 + 0$. Thus, the variance of total noise added in PFLR* is $2(d^2+d)(\Delta_f/\varepsilon_f)^2$. If we set $\varepsilon_f \geq (\Delta_f/\Delta_{\tilde{f}})\varepsilon$, PFLR* injects less noise.

4 EXPERIMENTS

4.1 Experiment Setup

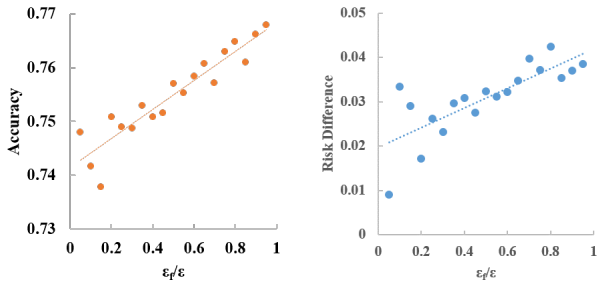
Dataset. We evaluate our methods on Adult [3] and Dutch [19]. For both datasets, we consider "Sex" as protected attribute and "Income" as decision. For unprotected attributes, we convert categorical attributes to one-hot vectors and normalize numerical attributes to $x \in [0, 1]$. The Adult dataset has 45222 records and 40 features. The Dutch dataset has 60420 records and 35 features.

Baselines. We compare the proposed differentially private and fair logistic regression models (**PFLR** and **PFLR***) with the following baselines: 1) a regular logistic regression model (**LR**); 2) a differentially private only LR using functional mechanism (**PrivLR**); 3) a fair only LR using Equation 11 as the fairness constraint (**FairLR**).

Metrics. We evaluate the performance of the proposed approaches and baselines on utility and fairness. We use *accuracy* as the utility metric and *risk difference (RD)* as the fairness metric. We run all models 10 times for each setting and report the mean and standard deviation of each metric.

4.2 Experimental Results

We first compare the performance of all five methods when $\varepsilon = 1$ ($\varepsilon_f = \varepsilon/2$ in PFLR*). As shown in Table 1, the regular logistic regression (LR) achieves the accuracy of 0.8380 on Adult and 0.8164 on Dutch, but it doesn't protect privacy nor achieves fairness (RD = 0.1577 and 0.1747, respectively). PrivLR has privacy protection but the accuracy decreases 11.42% on Adult and 17.52% on Dutch compared with LR as the result of the trade-off between privacy and utility. The risk difference of PrivLR is lower than LR, yet still larger than 0.05. The decrease of risk difference is mostly due to its low accuracy instead of fairness. FairLR achieves fairness (RD = 0.0095 on Adult and RD = 0.0299 on Dutch) as expected but it has no privacy guarantee. For PFLR and PFLR*, they both meet the privacy and fairness requirements. PFLR* has significantly higher accuracy than PFLR on both datasets (based on t-tests with p-values < 0.05). It indicates that PFLR* adds less noise to meet the same level of privacy guarantee.



(a) Accuracy **(b) Risk difference**

Figure 1: PFLR* with different privacy budget splits $\varepsilon_f/\varepsilon$ (Adult dataset, $\varepsilon = 10$)

Different Privacy Budgets. Table 2 shows how different settings of privacy budget ε affect our two methods and PrivLR. For PrivLR, its accuracy decreases dramatically when a stronger privacy requirement (smaller ε) is enforced. The risk difference of PrivLR decreases with the decrease of ε (the increase of noise). When ε is large, the accuracy is good and the risk difference is high. When ε is small, the accuracy is bad and the risk difference is low but with high variance.

For PFLR and PFLR*, when $\varepsilon = 0.1, 1$, PFLR* has significantly higher accuracy than PFLR on both Adult and Dutch (based on t-tests with p-values < 0.05). Especially, when $\varepsilon = 0.1$, PFLR's accuracy is only 0.6172 on Adult and 0.5069 on Dutch while PFLR*'s accuracy is 0.7491 on Adult and 0.6158 on Dutch. The accuracy of PFLR* is more consistent and relatively more resilient under different settings of privacy budget ε. When ε is relaxed to 100, PFLR and PFLR* have similar accuracy to FairLR (shown in Table 2). Overall, PFLR* outperforms PFLR especially when privacy budget is small.

Different Privacy Budget Splits $\varepsilon_f/\varepsilon$ for PFLR*. PFLR* splits the privacy budget ($\varepsilon = \varepsilon_f + \varepsilon_g$) into two parts: computing the fairness constraint (ε_g) and building the classification model (ε_f). Therefore, there is a trade-off between fairness and utility by controlling $\varepsilon_f/\varepsilon$. We further evaluate the performance of PFLR* in terms of accuracy and risk difference with various privacy budget splits by ranging the values of $\varepsilon_f/\varepsilon$ from 0.05 to 0.95 with an interval as 0.05. In Fig. 1a, we observe that with the increase of ε_f, the accuracy increases accordingly. This is because when ε_f keeps increasing, the privacy budget for the objective function becomes large, which reduces noise added to the classification model. For the risk difference, as shown in Fig. 1b, when ε_f increases, the risk difference increases. This is because PFLR* injects more noise to compute the fairness constraint. However, the risk difference is consistently smaller than 0.05 while increasing ε_f. Hence, as the result of a small sensitivity Δ_g, the utility of fairness constraint is well preserved even with a small ε_g.

5 CONCLUSION AND FUTURE WORK

In this work, we have developed two differentially private and fair logistic regression models, PFLR and PFLR*. PFLR is to apply the functional mechanism to the objective function with fairness

Table 1: Accuracy and risk difference (mean ± std.) of each method ($\varepsilon = 1$)

Method	Adult		Dutch	
	Accuracy	Risk Difference	Accuracy	Risk Difference
LR	0.8380 ± 0.0023	0.1577 ± 0.0064	0.8164 ± 0.0048	0.1747 ± 0.0033
PrivLR	0.7238 ± 0.0612	0.0502 ± 0.0581	0.6412 ± 0.0458	0.0739 ± 0.0574
FairLR	0.7739 ± 0.0521	0.0095 ± 0.0071	0.7673 ± 0.0064	0.0299 ± 0.0067
PFLR	0.7400 ± 0.0182	0.0213 ± 0.0258	0.6278 ± 0.0408	0.0206 ± 0.0204
PFLR*	0.7552 ± 0.0092	0.0053 ± 0.0070	0.6482 ± 0.0188	0.0430 ± 0.0265

Table 2: Accuracy and risk difference with different privacy budgets ε

	ε	PrivLR		PFLR		PFLR*	
		Accuracy	Risk Difference	Accuracy	Risk Difference	Accuracy	Risk Difference
Adult	0.1	0.6263 ± 0.1480	0.0883 ± 0.0805	0.6172 ± 0.1187	0.0351 ± 0.0493	0.7491 ± 0.0040	0.0028 ± 0.0039
	1	0.7238 ± 0.0612	0.0502 ± 0.0581	0.7400 ± 0.0182	0.0213 ± 0.0258	0.7552 ± 0.0092	0.0053 ± 0.0070
	10	0.7270 ± 0.0877	0.1459 ± 0.0798	0.7631 ± 0.0155	0.0338 ± 0.0255	0.7632 ± 0.0093	0.0204 ± 0.0140
	100	0.8295 ± 0.0032	0.1624 ± 0.0116	0.7835 ± 0.0318	0.0332 ± 0.0243	0.7913 ± 0.0200	0.0234 ± 0.0189
Dutch	0.1	0.5241 ± 0.0396	0.0317 ± 0.0187	0.5069± 0.0459	0.0441 ± 0.0245	0.6158 ± 0.0239	0.0516 ± 0.0204
	1	0.6412 ± 0.0458	0.0739 ± 0.0574	0.6278 ± 0.0408	0.0206 ± 0.0204	0.6482 ± 0.0188	0.0460 ± 0.0265
	10	0.7239 ± 0.0902	0.1346 ± 0.0563	0.7282 ± 0.0493	0.0211 ± 0.0152	0.7080 ± 0.0329	0.0220 ± 0.0208
	100	0.8154 ± 0.0042	0.1687 ± 0.0054	0.7681 ± 0.0054	0.0301 ± 0.0085	0.7618 ± 0.0144	0.0250 ± 0.0128

constraint as a penalty term. Our enhanced model, PFLR*, takes advantage of the connection between ways of achieving differential privacy and fairness and adds the Laplace noise with non-zero mean. The experimental results on two datasets demonstrate the effectiveness of two approaches and show the superiority of PFLR*. In this work, we consider logistic regression as the classification model and the covariance between decision boundary and the protected attribute as the fairness constraint. In future work, we plan to extend our methods to other classification models and other fairness constraints. Another research direction is to study allocation strategies of privacy budget, e.g., adding different amount of noise to coefficients containing different attributes.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1646654, 1502273 and 1523115.

REFERENCES

[1] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. 2011. Differentially Private Empirical Risk Minimization. *J. Mach. Learn. Res.* 12 (2011), 1069–1109.
[2] Bart Custers. 2013. Data Dilemmas in the Information Society: Introduction and Overview. In *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases*. 3–26.
[3] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[4] Cynthia Dwork. 2011. A Firm Foundation for Private Data Analysis. *Commun. ACM* 54, 1 (2011), 86–95.
[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science*. 214–226.
[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third*. 265–284.
[7] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In *Conference on Fairness, Accountability and Transparency*. 35–47.
[8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.
[9] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. Discrimination- and privacy-aware patterns. *Data Min. Knowl. Discov.* 29, 6 (2015), 1733–1782.
[10] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.
[11] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *ICDMW*.
[12] Ugo Pagallo. 2012. On the Principle of Privacy by Design and its Limits: Technology, Ethics and the Rule of Law. In *European Data Protection: In Good Health?* 331–346.
[13] Walter Rudin. 1953. *Principles of mathematical analysis*. McGraw-Hill.
[14] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On Convexity and Bounds of Fairness-aware Classification. In *WWW*.
[15] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.
[16] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. 2012. Functional mechanism: regression analysis under differential privacy. *PVLDB* 5, 11 (2012), 1364–1375.
[17] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI*.
[18] Lu Zhang, Yongkai Wu, and Xintao Wu. 2018. Achieving Non-Discrimination in Prediction. In *IJCAI*.
[19] I. Žliobaite, F. Kamiran, and T. Calders. 2011. Handling Conditional Discrimination. In *ICDM*.

Augmenting Intelligence with Bias-aware Humans-in-the-loop (HumBL2019) Chairs' Welcome

It is our great pleasure to welcome you to the *WWW 2019 Augmenting Intelligence with Bias-aware Humans-in-the-loop (HumBL2019)*. The goal of this workshop is to bring together researchers and practitioners in various areas of AI (i.e., Machine Learning, NLP, Computational Advertising, etc.) and crowdsourcing, to explore new pathways of the human-in-the-loop paradigm and disentangle the relationship between crowdsourcing and bias management. We aim to discuss methods where the user has the ability to heavily influence the outcome of the process by providing feedback to the system, as well as the opportunity to grab different perspectives about the underlying domain and understand the step by step machine process leading to a certain outcome. Additionally, we focus on one of the major concerns in AI research: *being able to explain and understand the results as well as identifying, avoiding and/or mitigating bias in the underlying data that might lead to unfair or unethical conclusions*. Bringing humans and machines together creates a natural symbiosis for accurate interpretation of data at scale.

We focus on the partnership between humans and machines in social computing systems, with a special emphasis on *human factors and ethics of crowdsourcing techniques*, and a wide-range of *domain applications*, such as healthcare, NLP technologies, translation, data quality control and sentiment analysis. As human-in-the-loop processes are intrinsically related to *machine learning*, we consider solutions that deal with sparse and noisy annotated data, as well as workflows with crowd-supported active learning. Additionally, we address challenges in the *crowdsourced data collection* process, including task design, collecting data while respecting privacy constraints and dealing with multi-linguality. Finally, we discuss the different types of *bias in crowdsourcing* (e.g., gender-related, cultural, cognitive, experience and knowledge bias) that can occur in different steps of the crowdsourcing process, from recruiting and task selection to result aggregation, and analyse methods to monitor and mitigate biases in algorithmic systems through crowdsourcing, as it can help diversify sources of information.

Below we summarize the accepted contributions presented at the workshop. You can find the full workshop program at the workshop website[1]. This year, for the first time, the HumBL workshop ran a joint keynote series in collaboration with the SAD2019 workshop[2] featuring five keynotes from industry and academia: Maria Stone from Apple, Brad Klingenberg from StitchFix both presented at SAD2019, and Anima Anandkumar from NVIDIA, Anurag Batra from Google AI and Jon Chamberlain from the University of Essex presented at HumBL2019.

Three full and five short papers were presented at HumBL2019 workshop. They covered a wide range of topics related to the efficient and effective combination of both machine and crowd computation. In overall, the workshop presents a nicely diverse set of use cases and domains where these topics were studied. Empirical results were discussed with respect to (1) methods for data quality assurance and

[1] https://humlworkshop.github.io/HumBL-WWW2019/

[2] https://sadworkshop.wordpress.com/keynotes/

labeling task efficiency, (2) the role of gamification elements for improving crowd performance as well as the role of quantum mathematics to simulate human behavior. Following, an overview of the papers:

- *Crowdsourcing with Fairness, Diversity and Budget Constraints*, by Goel et al, presents a novel task assignment algorithm that maximizes the expected accuracy of crowdsourced data and at the same time ensures that the errors in the data are as fair as possible, minimizing the bias.
- *Debiasing Collaborative Filtering Recommender Systems*, by Sun et al, tackles the problem of introducing bias in recommender systems due to feedback loops and proposes various mechanisms to debias the recommender systems.
- *Uncovering Bias in Ad Feedback Data: Analyses & Applications*, by Marc Bron, Ke Zhou, Andrew Haines and Mounia Lalmas, presents initial insights into sources and types of bias that can be identified in ads feedback and proposes a variety of models to correct them.
- *Identifying High Value Opportunities for Human in the Loop Lexicon Expansion*, by Alba et al, proposes a measure that given a set of lexicons, helps identify the lexicon that would profit the most from a human-in-the-loop lexicon extension.
- *Task Oriented Data Exploration with Human-in-the-Loop*, by Alba et al, presents a method to semantically annotate nodes and clusters of nodes in a data center based on occurrying logged activity with humans-in-the-loop, that aids in data center migration planning.
- *Implicit Bias in Crowdsourced Knowledge Graphs*, by Demartini, highlights the risks of introducing bias in knowledge graphs created by means of crowdsourcing and proposes methods to trace bias lineage to create awareness in data consumers.
- *Towards Natural Language Search of Time Series Data*, by Imani et al, presents a human-in-the-loop method to facilitate the search of multidimensional time series through the use of natural language tokens.
- *I'm Lonely. Who should I talk to?*, by Fratamico presents a human-in-the-loop approach to mediate and facilitate online interactions among people by highlighting their common interests.

Lora Aroyo
Google

Alessandro Checco
The University of Sheffield

Gianluca Demartini
The University of Queensland

Ujwal Gadiraju
L3S Research Center

Anna Lisa Gentile
IBM Almaden Research

Oana Inel
Delft University of Technology

Cristina Sarasua
University of Zurich

Augmenting Intelligence with Bias-aware Humans-in-the-loop (HumBL2019) Organization

Organization Committee:
Lora Aroyo *(Google, USA)*
Alessandro Checco *(The University of Sheffield, UK)*
Gianluca Demartini (*The University of Queensland, Australia)*
Ujwal Gadiraju *(L3S Research Center, Germany)*
Anna Lisa Gentile *(IBM Almaden Research Center, USA)*
Oana Inel *(Delft University of Technology, The Netherlands)*
Cristina Sarasua *(University of Zurich, Switzerland)*

Program Committee:
Alessandro Bozzon *(Delft University of Technology, The Netherlands)*
Irene Celino *(CEFRIEL, Switzerland)*
Lydia Chilton *(Columbia University, USA)*
Djellel Difallah *(NYU, USA)*
Carsten Eickhoff *(Brown University, USA)*
Anca Dumitrache *(Vrije Universiteit Amsterdam, The Netherlands)*
Daniel Gruhl (*IBM Almaden Research Center, USA)*
Ricardo Kawase *(mobile.de GmbH / eBay Inc.)*
Rochelle Laplante *(Independent)*
Walter Lasecki *(University of Michigan, USA)*
Praveen Paritosh *(Google)*
Bibek Paudel *(University of Zurich, Switzerland)*
Marta Sabou *(Vienna University of Technology, Austria)*
Mike Schaekermann *(University of Waterloo, USA)*
Elena Simperl *(University of Southampton, UK)*
Amrapali Zaveri *(Maastricht University, The Netherlands)*

Are Two Heads Better Than One? An Exploration of Ambiguity in Crowd-Collected Language Decisions from the Phrase Detectives Game.

Jon Chamberlain
School of Computer Science and Electronic Engineering
University of Essex, Colchester, Essex UK, jchamb@essex.ac.uk

ABSTRACT

The online game-with-a-purpose Phrase Detectives (https://www.phrasedetectives.com) [1] has been collecting decisions about anaphoric coreference in human language for over 10 years (4 million judgements from 40,000 players). The game was originally designed to collect multiple valid solutions for a single task, which complicated aggregation but created a very rich (and noisy) dataset [2]. Analysis of the ambiguous player decisions highlights the need for understanding and resolving disagreement that is inherent in language interpretation. This talk will present some of the interesting cases of ambiguity found by the players of Phrase Detectives (a dataset that will be made available to the research community later this year [3]) and discuss the statistical methods we have been working on to harness crowds that disagree with each other [4, 5].

ACM Reference format:

Jon Chamberlain. 2019. Are Two Heads Better Than One? An Exploration of Ambiguity in Crowd-Collected Language Decisions from the Phrase Detectives Game.. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317311

1 Speaker biography

Dr Jon Chamberlain is a lecturer and applied research scientist based at the University of Essex, England. He has been the lead developer behind the Phrase Detectives game-with-a-purpose for over 10 years and is co-investigator on the ERC Advanced Grant Disagreements and Language Interpretation (DALI) project that builds on early work to understand ambiguity in human language.

2 ACKNOWLEDGMENTS

The speaker is funded by an ERC Advanced Grant – ERC-2015-AdvG.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317311

REFERENCES

[1] Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Transactions on Interactive Intelligent Systems, 3(1):1–44.

[2] Chamberlain, J., Poesio, M., and Kruschwitz, U. (2016). Phrase Detectives corpus 1.0: Crowdsourced anaphoric coreference. In Proceedings of LREC, Portoroz, Slovenia.

[3] Poesio, M., Chamberlain, J., Paun, S., Kruschwitz, U., and Yu, J. (2019 forthcoming). "A Crowdsourced Corpus of Multiple Judgments and Disagreement on Anaphoric Interpretation." In Proceedings of The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis. Association for Computational Linguistics.

[4] Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., and Poesio, M. (2018). "Comparing bayesian models of annotation." Transactions of the Association for Computational Linguistics.

[5] Paun, S., Chamberlain, J., Kruschwitz, U., Yu, J., and Poesio, M. (2018). "A probabilistic annotation model for crowdsourcing coreference." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1926–1937, Brussels, Belgium. Association for Computational Linguistics.

Identifying High Value Opportunities for Human in the Loop Lexicon Expansion

Alfredo Alba
IBM Research Almaden, CA, US
aalba@us.ibm.com

Chad DeLuca
IBM Research Almaden, CA, US
delucac@us.ibm.com

Anna Lisa Gentile
IBM Research Almaden, CA, US
annalisa.gentile@ibm.com

Daniel Gruhl
IBM Research Almaden, CA, US
dgruhl@us.ibm.com

Linda Kato
IBM Research Almaden, CA, US
kato@us.ibm.com

Chris Kau
IBM Research Almaden, CA, US
ckau@us.ibm.com

Petar Ristoski
IBM Research Almaden, CA, US
petar.ristoski@ibm.com

Steve Welch
IBM Research Almaden, CA, US
welchs@us.ibm.com

ABSTRACT

Many real world analytics problems examine multiple entities or classes that may appear in a corpus. For example, in a customer satisfaction survey analysis there are over 60 categories of (somewhat overlapping) concerns. Each of these is backed by a lexicon of terminology associated with the concern (e.g., “Easy, user friendly process" or "Process confusing, too many handoffs”). These categories need to be expanded by a subject matter expert as the terminology is not always straight forward (e.g., “handoffs” may also include “ping-pong” and “hot potato” as relevant terms).

But given that Subject Matter Expert time is costly, which of the 60+ lexicons should we expand first? We propose a metric for evaluating an existing set of lexicons and providing guidance on which are likely to benefit most from human-in-the-loop expansion. Using our ranking results we achieved $\approx 4\times$ improvement in impact when expanding the first few lexicons off our suggested list as compared to a random selection.

ACM Reference Format:
Alfredo Alba, Chad DeLuca, Anna Lisa Gentile, Daniel Gruhl, Linda Kato, Chris Kau, Petar Ristoski, and Steve Welch. 2019. Identifying High Value Opportunities for Human in the Loop Lexicon Expansion. In *HumBL2019. The third international workshop on Augmenting Intelligence with Bias-Aware Humans-in-the–Loop. In the Web Conference 2019 Companion volume.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317305

1 INTRODUCTION

The data in real world text analytics problems is messy. It typically contains a plethora of categories with varying degrees of relatedness towards the task at hand. Each category can be represented by a lexicon; containing a representation, as exhaustive as possible, language variants authors use to express an entity belonging to the given category. With the proper set of lexicons that represent the necessary categories, further analysis can be conducted, such as survey classification, semantics, category relationship analysis, etc. These analyses extract the insights that can ultimately enable a data driven business decision making process.

We illustrate this with a case study of customer service satisfaction surveys. These surveys are typically designed to collect data to answer specific business questions, e.g. to understand customers' concerns about specific categories of issues. More than one category may appear in a single survey; e.g., a customer can be satisfied with the service agent's empathy towards them while still being upset about the amount of time it took to solve the issue. In order to derive the desired insights from the data, it is essential to identify and effectively extract the meaningful category identifiers.

As these are surveys written in the customer's own language, the various phrasings, misspellings, etc. need to be interpreted and well understood. Acquiring a deep understanding of the language a specific community uses to describe many categories of issues - in this specific case more than 60 - at any given point in time is a considerable challenge. A brute force approach for this task can be effective, but extremely time consuming. Additionally, asking a Subject Matter Expert (SME) to quickly score tens of thousands of documents becomes an daunting task as review fatigue sets in.

In our approach, the SMEs quickly curate the lexicons for each category from their experience and they manually inspect as many survey examples as possible. While the manual approach can be effective, it provides only a point in time solution. Over time, the language of a community drifts due to both internal changes (e.g., call centers moving, new product releases, etc.) as well as external changes in the common usage (slower but still of importance over decades of data). Thus, lexicon curation becomes a continuous improvement exercise to maintain it's relevancy.

Humans, even SMEs, are not perfect in their performance, as demonstrated by the number of human reliability assessment techniques [15]. The factors that impact the human reliability are numerous [16, 25]. These factors have a considerable impact on lexicons solely curated by an SME using a manual process. Therefore the higher the target quality of the manually curated data, the higher the cost (1).

On the other hand, traditional distantly or closely supervised approaches require considerable up front investments in developing

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
HumBL2019, May 2019, San Francisco, CA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317305

Figure 1: In real world analytics applications, the Subject Matter Expert is often confronted with dozens of semantic assets such as lexicons that must be refreshed. Without guidance, they can end up spending time working on lexicons whose improvement has very little impact on the overall business objectives.

large amounts of high quality representative training data; and it is subject to the same decay over time as highlighted above.

This might lead one to consider an unsupervised approach. However with these the cost can be even higher due to all the data scientist time that has to be spent on feature engineering, which is indispensable in assuring there are no biases or poisoned features, potentially rendering the resulting model unusable [5, 6]. Although recent work [4] has taken aim at addressing this issue via adversarial networks, among other approaches, depending on the target model this remains a very active area of research without a definitive solution.

An alternative is a human-in-the-loop system where the human is assisted by a data driven cognitive system in the task of lexicon curation. The system helps by providing candidates for human adjudication. As the candidates are adjudicated the system learns the semantics of the categories at hand and rapidly produces higher quality candidates for the SME to adjudicate.

This human-in-the-loop system maximizes the impact of the SME's time and effort. By putting the greatest value on the human's time, resources usage is optimized in a way that reflects real world requirements. Business value is achieved by identifying the most efficient path, given available knowledge and resources.

The system provides the added benefit of keeping the focus on areas of rapid change. It might be common for lexicons in most categories to rarely require modification, however some categories undergoing rapid change will surface more frequently, thus exposing themselves as more dynamic.

We propose a novel technique to proactively identify which lexicons can benefit most from expansion. This can be used both to order the expansion efforts, as well as (using a threshold) monitor a collection of lexicons and alert when expansion is warranted.

The major contribution of this paper is a set of novel features and way of combining them which can be used to detect when expansion is likely to be beneficial. We discovered that the most important features to take into account are: the size of the initial lexicon, the number of initial hits, the lexicon intra-similarity and the confidence of the used annotation model.

2 STATE OF THE ART

Lexicons, ontologies, and linguistic resources are the backbone of many NLP and information retrieval systems and applications. The automatic construction of such resources has been the focus of research for many years. Riloff and Jones [20] propose one of the first approaches to extract dictionaries from unstructured text. Starting from a few seed terms, it learns extraction patterns, which are then used to expand the seeds and iteratively repeat the process. In the following years, a number of similar approaches have been developed [3, 7, 14, 21]. However, all these approaches require NLP parsing for feature extraction, and thus have a reliance on syntactic information for identifying quality patterns. Hence, such approaches underperform on not-so-well structured or or grammatically incorrect text content, such as user-generated text. Furthermore, completely automatic iterative methods without human-in-the-loop can easily generate semantic drift - a few spurious extraction patterns can exponentially increase the inclusion of incorrect items in the lexicon.

To address these issues, recent works propose feature-antagonistic approaches for dictionary generation [2, 12]. These approaches couple deep neural language models with tight human supervision to assist the user in building and maintaining domain-specific dictionaries. While these approaches are able to efficiently and effectively extend lexicons, in cases where the number of input lexicons is high, they do not provide guidance on which lexicons are likely to benefit most from human-in-the-loop expansion.

While there are no approaches in the literature that address this problem directly, there are many approaches that, given a set of items, can be used for ranking and selecting the most promising item that will lead to the highest reward. For example, active learning tries to identify the most important data instances to be labeled by a human oracle, i.e., identifies the data points that would improve the model performance the most. In the literature there are many ranking and selection strategies proposed [24], e.g., uncertainty sampling [18], density weighted uncertainty sampling [10, 19], diversity [8], QUIRE [13] and Bayesian methods such as BALD (Bayesian Active Learning by Disagreement) [11].

Another type of approaches that focus on the problem of ranking and selecting the most beneficial items, are bandit approaches [22, 23]. More precisely, bandit approaches develop a set of strategies that given a set of choices/items are able to rank the probability for selecting each of them in a way that maximizes the expected gain. This problem has been studied for many years resulting in a plethora of existing approaches [17].

While there are many approaches for document ranking and selection in machine learning, to the best of our knowledge, no existing approach addresses the challenges of ranking lexicons and providing guidance on which are likely to benefit most from human-in-the-loop expansion.

Table 1: Surveys corpus and lexicons statistics. Average (avg.), minimum (min), maximum (max) and median (med.) values for the lexicons size, number of Total Tagged Documents (TTD), and intra-lexicon similarity, in the original lexicons (Original) and after the extension of the lexicons (Extended)

	#Surveys	#Lex.	Lexicon Size				#TTD					Intra-Lexicon Similarity			
			avg.	min	max	med	total	avg.	min	max	med.	avg.	min	max	med
Original	378,000	65	18.00	2	96	8	147,714	5,908.56	33	44,925	1,252	0.39	0.13	0.86	0.39
Extended	378,000	65	51.96	5	125	45	203,839	8,153.56	73	48,965	2,543	0.35	0.24	0.52	0.34

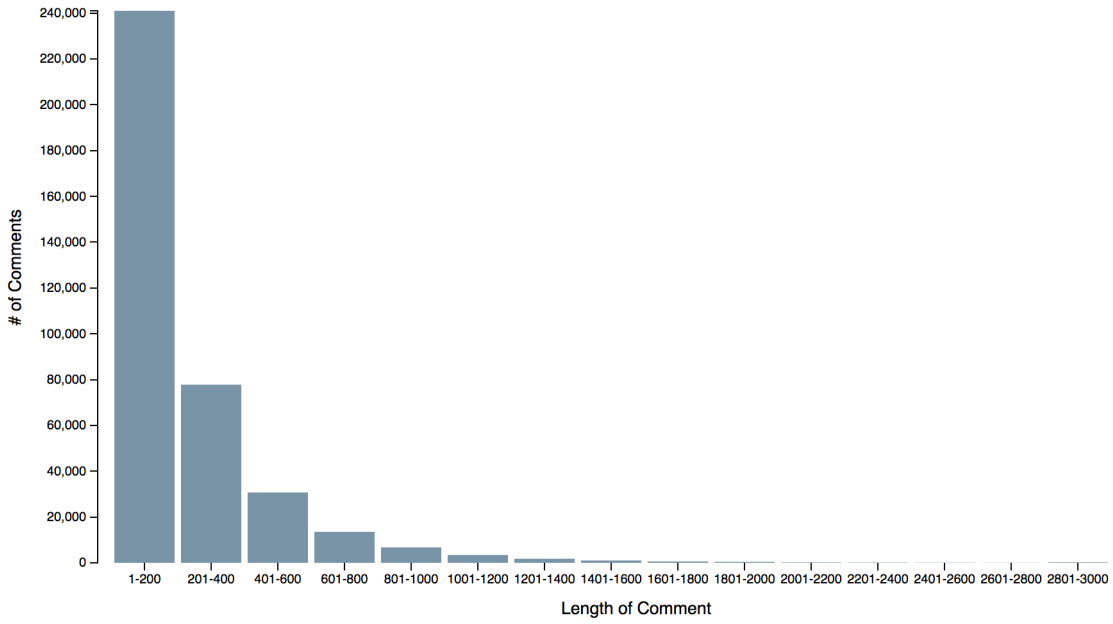


Figure 2: Customer Satisfaction surveys contain a free text field as well as structured questions. The free text responses tend to be very short, but there are cases of respondents including upwards of a page of text detailing their responses on their interactions with the call center.

3 APPROACH

The measure we are seeking to optimize is the count of "Total Tagged Documents" (*TTD*) in the corpus we are examining. Thus, if adding the phrase "great customer service" resulted in 2,900 more documents being tagged (as they each had this phrase appearing), we would say that adding that phrase had a "score" of +2, 900.

Phrased as such, we have a classic prediction problem ahead of us. We wish to produce an ordered list of lexicons to expand, such that the first lexicon we suggest is most likely to result in the highest increase in *TTD*. We score ourselves using the Spearman rank correlation metric for list ordering.

There are then two steps; identifying (and developing) features that correlate meaningfully with *TTD*, followed by identifying an effective way of combining those features to create a useful ordering.

3.1 Features

Features available to us are those that can be measured on the a priori lexicons before any human involvement. We note that in the case of "continual improvement" situations, there may be more data available, but we leave that for future work.

Several features were found to be useful. Two basic ones were number of terms in the lexicon, and the *TTD* of the lexicon initially. The intuition here is that large lexicons are perhaps more fully expanded already, and that lexicons that cover a large number of documents are focusing on concepts that are prevalent in the source corpus.

Another feature was obtained via running a `word2vec` style expander (Explore and Exploit [1, 12]) and looking at how "close" in vector space the terms in the initial lexicon were. This gives a sense of how "tight" the initial terminology was. We expect a more fully expanded lexicon to be more "diffused" (having less coherence) as more "edge concepts" are included. The feature is calculated as the average cosine similarity of all the terms in the lexicons (the higher the value, the more similar are the terms inside the lexicon).

A fourth feature type was obtained by running an initial Explore and Exploit expansion to look at the curve of confidence and cumulative confidence per term for the first 20 terms. This aims to be

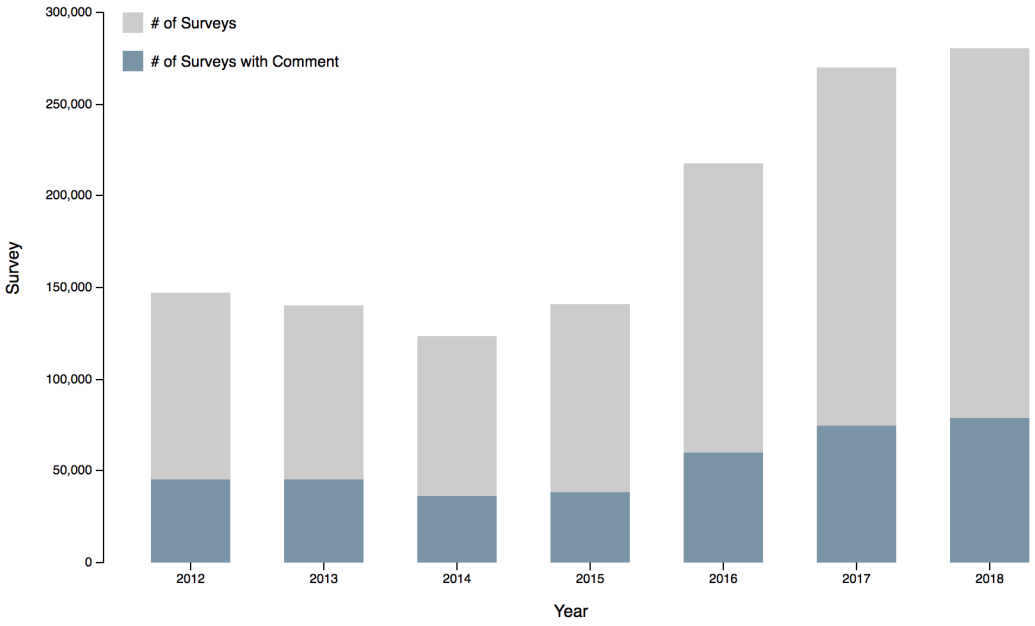


Figure 3: Surveys span 7 years. Even though most do not contain a free text comment, around 50,000 a year do – many more than it would be practical to human review for each of 60+ categories.

a measure of the "low hanging fruit" of the algorithm. While it is not exact (as there has been no human involvement, nor is it very much like the Glimpse[9] engine which is also used by the human) it does prove to be one of the better single features.

Our fifth feature looks at the number of hits a lexicon gets each year for the 7 years we have available in the survey dataset. We normalize by the total number of surveys in the year and then fit a line to the results. This "slope" gives us a measure of how quickly the initial terms are decreasing in prevalence, with the notion this can be a measure of how quickly the terms are going "stale" as language evolves around the entity of interest.[1]

3.2 Regression Models

We approach the problem of ranking lexicons as a regression problem, i.e., we try to predict the increase in *TTD* after the lexicon expansion, which is then used to rank the lexicons. To effectively combine the extracted features, we use 4 standard regression models: (i) Linear Regression, (ii) K-Nearest Neighbors, (iii) Decision Trees and (iv) Neural Network.

4 EVALUATION

4.1 Dataset

The dataset used for this experiment consists of customer survey results spanning 7 years. The survey asks the customer to rate their experience in a number of categories using a numerical scale.

[1]Note: While this applies only to the single dataset we tested on, the approach is transferable to any "tune up" of a domain. This is especially true when you are doing year-to-year tune-ups and have data to fit a predictor to, based on what worked well in the last round of tune-ups.

Table 2: Spearman's rank correlation ρ for lexicon rank prediction, using 4 models: Linear Regression (LR), K-Nearest Neighbors (KNN), Neural Network (NN) and Decision Tree (DT)

Model	ρ
LR	0.32
KNN	0.71
NN	0.41
DT	**0.86**

There is also an optional unstructured text field for the customer to add any additional comments. All questions are optional, so many questions are left unanswered. The dataset contains survey results from 2102 to 2018. There are a total of 132 million surveys in the dataset, about 378K of which contain free form text comments. The length of the comments typically varies between a single word and a couple of paragraphs of text. The distribution of the length of the comments is shown in Figure 2. Figure 3 shows the distribution of comments per year. We can observe that there is a slight increase of comments in the last 3 years. We use the set of 378K surveys to build the Explore and Exploit model [2], which is used for extending the lexicons. We used a set of 65 lexicons which were expanded by a group of 8 Subject Matter Experts. The statistics of the used lexicons before and after the expansion are shown in Table 1. Given an input text corpus and a set of seed examples, the Explore and Exploit approach first builds a neural language model on the input text corpus, and then runs in two phases to identify new potential lexicon entries. The Explore phase tries to identify similar instances

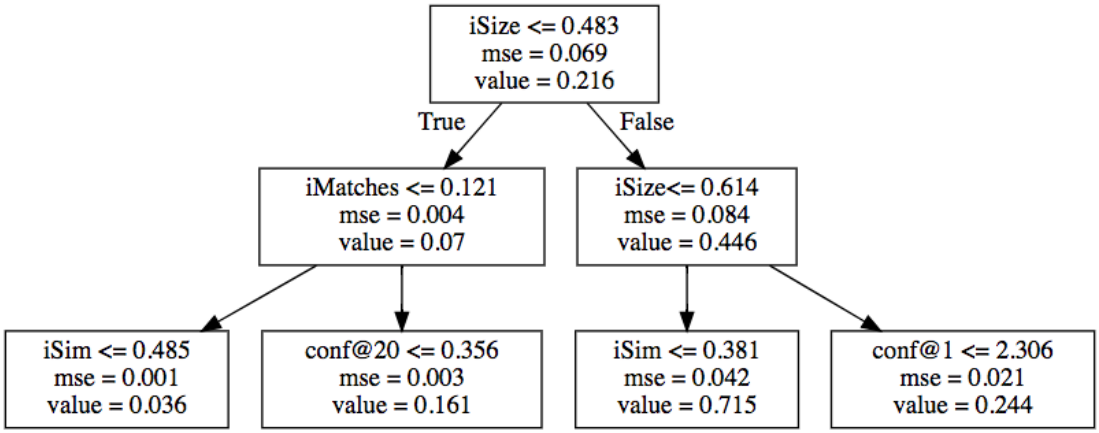


Figure 4: Decision tree model visualization.

to the dictionary entries that are present in the input text corpus, using term vectors from the neural language model to calculate a similarity score. The Exploit phase tries to construct more complex multi-term phrases based on the instances already in the input dictionary. The SME is closely involved in the processed to avoid semantic drift and incorrect entries in the lexicon.

4.2 Results

We consider 4 regression models:

- Linear Regression (LR)
- K-Nearest Neighbors (KNN), with K=3
- Neural Network (NN) with one input layer, three fully connected hidden layers with ReLU activation function, with 10, 8 and 6 neurons respectively, and an output layer with linear activation function.
- Decision Tree (DT), using mean squared error for node splitting criteria and with a maximum tree depth of 5.

For the implementation of the models we use the Python *sklearn* library.

To evaluate the models, we use the Spearman's rank correlation as an evaluation metric, i.e., we evaluate the correlation of the lexicon rank produced by the models and the real rank of the lexicons.

The results are show in Table 2. The Spearman's rank correlation is highest when using a Decision Tree model. Closer analysis of the decision tree structure reveals that the most important features are the size of the initial lexicon (iSize), the number of initial hits (iMatches), the lexicon intra-similarity (iSim) and the model confidence at 1 (conf@1) and at 20 (conf@20). This means that the smallest lexicons with high coherency (e.g. the variety of terms is very low) need to be extended first. The top 3 layers of the decision tree model are shown in Figure 4. We can see that the neural network model is not able to learn a good representation of the data, as the dataset is rather small. The K-Nearest Neighbors model performs quite well with both sets of features, while the Linear Regression could not learn a good model.

We compare the final model to a random selection from the lexicons to expand by calculating *TDD* 3 and 5. Compared to a random selection at 3, we see 406% uplift in *TDD*, and at 5 we still see 387% uplift. This represents a significant time savings/increase in impact for the human Subject Matter Experts (Figure 5).

5 CONCLUSIONS AND FUTURE WORK

In any human-in-the-loop task, human time is by far the most constrained resource a system has to work with. As human-in-the-loop rolls out to more niche domains, there is a limit to how long human involvement can be justified - hence the importance of making the most out of it.

For large analytics projects with dozens of lexicons, keeping them up to date is a time consuming but necessary task. Our approach meshes well with the real world "quarterly update" notion of lexicon maintenance – by identifying where human involvement will make the most difference.

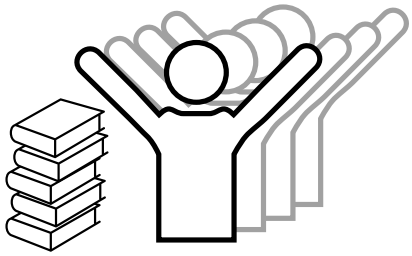


Figure 5: By providing Subject Matter Experts with an ordered list of lexicons, they can start with the most impactful first. The head of this list is 4× more impactful than a random selection.

By allowing the human to identify the lexicons that are most likely to benefit from expansion, time can be planned, experts marshaled, and improvements made in an incremental manner, with the very high likelihood ($\rho = 0.86$) that their work will make the most difference. Overall, this results in a 4× increase in impact for human Subject Matter Experts who are expanding just a few lexicons.

While this approach has worked well for this case, it would be good to test the approach on a wider variety of lexicon sets. One

of especial interest is the Wikipedia/DBpedia categories - can our approach identify new entities to tag into a category, or perhaps even new articles that should be authored?

Additionally, in an incremental setting (e.g., identify a good lexicon to expand, expand it, repeat), taking advantage of the accuracy of the prediction to dynamically adjust the splits in the tree may help better identify the next lexicon to enrich.

REFERENCES

[1] Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. 2017. Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop. In *Proc. of the Knowledge Capture Conference*. ACM, 24.

[2] Alfredo Alba, Daniel Gruhl, Petar Ristoski, and Steve Welch. 2018. Interactive Dictionary Expansion using Neural Language Models. (2018).

[3] Rie K Ando. 2004. *Semantic lexicon construction: Learning from unlabeled data via spectral analysis*. Technical Report. IBM THOMAS J WATSON RESEARCH CENTER YORKTOWN HEIGHTS NY.

[4] Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, and Jaehoon Amir Safavi. 2017. Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 103–110.

[5] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. 2006. Can machine learning be secure?. In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. ACM, 16–25.

[6] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389* (2012).

[7] Sebastian Blohm and Philipp Cimiano. 2007. Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction. In *PKDD 2007*. Springer, 18–29. DOI:http://dx.doi.org/10.1007/978-3-540-74976-9_6

[8] Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*.

[9] Anni Coden, Daniel Gruhl, Neal Lewis, Michael Tanenblatt, and Joe Terdiman. 2012. SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora. *Proc. - 2012 IEEE 2nd Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2012* (2012), 33–39.

[10] Pinar Donmez, Jaime Carbonell, and Paul Bennett. 2007. Dual strategy active learning. In *ECML*. Springer.

[11] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *ICML*.

[12] Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. 2019. Explore and Exploit. Dictionary Expansion with Human-in-the-Loop. In *European Semantic Web Conference (TO APPEAR)*. Springer International Publishing.

[13] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. 2010. Active learning by querying informative and representative examples. In *NIPS*.

[14] Sean P Igo and Ellen Riloff. 2009. Corpus-based semantic lexicon induction with web-based corroboration. In *Proc. of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*. Association for Computational Linguistics, 18–26.

[15] Barry Kirwan. 1992. Human error identification in human reliability assessment. Part 2: detailed comparison of techniques. *Applied ergonomics* 23, 6 (1992), 371–381.

[16] Barry Kirwan. 1996. The validation of three Human Reliability Quantification techniquesÃćâĆňâĂİTHERP, HEART and JHEDI: Part 1ÃćâĆňâĂİtechnique descriptions and validation issues. *Applied ergonomics* 27, 6 (1996), 359–373.

[17] Volodymyr Kuleshov and Doina Precup. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028* (2014).

[18] David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *ICML*.

[19] Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *ICML*. ACM.

[20] Ellen Riloff, Rosie Jones, and others. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*. 474–479.

[21] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *Proc. of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. 25–32. DOI:http://dx.doi.org/10.3115/1119176.1119180

[22] Herbert Robbins. 1985. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*. Springer, 169–177.

[23] Gunter Rudolph. 1997. Reflections on bandit problems and selection methods in uncertain environments. In *International Conference on Genetic Algorithms*. Citeseer.

[24] Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114.

[25] JC Williams. 1988. A data-based method for assessing and reducing human error to improve operational performance. In *Human Factors and Power Plants, 1988., Conference Record for 1988 IEEE Fourth Conference on*. IEEE, 436–450.

Task Oriented Data Exploration with Human-in-the-Loop. A Data Center Migration Use Case.

Alfredo Alba, Chad DeLuca, Anna Lisa Gentile, Daniel Gruhl,
Linda Kato, Chris Kau, Petar Ristoski, Steve Welch
aalba@us.ibm.com,delucac@us.ibm.com,annalisa.gentile@ibm.com,dgruhl@us.ibm.com,
kato@us.ibm.com,ckau@us.ibm.com,petar.ristoski@ibm.com,welchs@us.ibm.com
IBM Research Almaden, CA, US

ABSTRACT

Data exploration is a task that inherently requires high human interaction. The subject matter expert looks at the data to identify a hypothesis, potential questions, and where to look for answers in the data. Virtually all data exploration scenarios can benefit from a tight human-in-the-loop paradigm, where data can be visualized and reshaped, but also augmented with missing semantic information - that the subject matter expert can supplement *in itinere*. In this demo we show a novel graph-based data exploration model where the subject matter expert can annotate and maneuver the data to answer specific questions. This demo specifically focuses on the task of migrating data centers, logically and/or physically, where the subject matter expert needs to identify the function of each node - a server, a virtual machine, a printer, etc - in the data center, which is not necessarily directly available in the data and to be able to plan a safe switch-off and relocation of a cluster of nodes. We show how the novel human-in-the-loop data exploration and enrichment paradigm helps designing the data center migration plan.

ACM Reference Format:
Alfredo Alba, Chad DeLuca, Anna Lisa Gentile, Daniel Gruhl, Linda Kato, Chris Kau, Petar Ristoski, Steve Welch. 2019. Task Oriented Data Exploration with Human-in-the-Loop. A Data Center Migration Use Case.. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3317306

1 INTRODUCTION

Understanding the functions implemented within a data center is an extremely challenging problem, due to countless machine re-configuration, software updates, changing software installation, failures, malevolent external attacks, etc. Being able to quickly and precisely characterize the nature, role, connections - which are often not explicitly declared - of the multitude of nodes in a data center becomes paramount when planning data center migrations.

Migrating data centers - either physically or moving applications to the cloud - is a time and resource intensive task. Preparing a migration plan involves intensive data analysis, often based on the logs and the network activities of each node in the data center. Discovering and understanding connections and dependencies can be very laborious and missing any of them can result in unplanned failures during the migration. Traditional data analysis tools offer little support during the plan-making phase, which can take many man hours.

In this work we propose a data exploration solution that allows the subject matter expert to interactively augment the data with structured knowledge and semantic information which is not initially present in the data. We combine traditional Information Extraction techniques together with human-in-the-loop learning to construct a semantic representation of the functions provided by the data center.

The contribution of this work is twofold. First, we propose a novel technique to extract semantic knowledge about nodes or clusters of nodes in the data center. While available structured knowledge about data center nodes - processes numbers, port numbers, IP addresses... - is readily available, semantic knowledge about each node is not formally encoded. We propose the use of logs from nodes in the data center to extract semantic information about the running processes. The knowledge extraction is performed with a human-in-the-loop model: we identify repeating patterns in the logs and ask a subject matter expert to label them (e.g. a certain log might indicate that the node hosts a "database"), we then generate regular expressions to label similar processes (i.e. similar logs) accordingly. The operation is repeated iteratively until the subject matter expert is satisfied with the label coverage. Second we leverage the added semantic knowledge, together with all other already available information, within a visual discovery framework to support data center migration planning.

In this demo we will show how to explore the functions of a data center using our novel exploration tool. We will demonstrate the value provided by the semantic knowledge added interactively. During the live demo we will showcase the use of the tool to answer questions such as "Find all nodes running a database application" followed by "Find all webservers connecting to those databases".

The main advantage of our solution is that it enables subject matter experts to quickly explore, characterize and augment complex data. Specifically in this case information about data centers, where we help the subject matter experts to quickly combine (i) information in the nodes' logs together with (ii) iteratively added knowledge that is not available from the logs to create a cartography of the data center.

In the following we will give an account of available research on data exploration, specifically focusing on data center migrations (Section 2) and then describe the specific use case that will be

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317306

demonstrated during the live demo (Section 3). In Section 4 we depict conclusions of this work and our planned future work.

2 STATE OF THE ART

The Knowledge Discovery pipeline proposed by Fayyad et al. [7] has been widely accepted and most of the existing data analysis systems comply to its principles. Knowledge discovery is the process of identifying unsuspected relationships and summarizing the data in novel ways that are both understandable and useful to the data owner [9]. The knowledge discovery process usually comprises five steps, i.e. data selection, data processing, data transformation, data analysis, and evaluation and interpretation of the discoveries.

In the case of knowledge discovery for data center exploration, the most abundant type of available data is in the form of machine logs. Therefore, a great deal of work has been devoted to adapt the steps of the pipeline to this specific format. The common ground for knowledge discovery from log-like data is the usage of rules - in the form of regular expressions, filters, etc. Lemoudden and El Ouahidi [11] use tokenization, regular expressions, dictionaries and timestamp filter. Similarly, we use regular expressions and dictionaries, but rather than using a pre-constructed set of filters, we iteratively construct those during data exploration with a human-in-the-loop model. The timestamp is a feature that is also used in many other works analyzing logs [2, 10]. Nonetheless, these works are focused on the identification of user sessions, e.g., in Web based search [10] or within e-commerce frameworks [2]. While the temporal component is paramount for certain analysis tasks (e.g. discovering cyber-attacks in data centers), it is not a key component to depict a cartography of the data center, i.e., for understanding the functionalities provided by each node (or clusters of nodes) - and in this scenario we find that helping the user to semantically tag the nodes leads to better insights. In other words, understanding the nature and context of a particular event is far more valuable to our use case than understanding exactly when that event occurred.

In terms of generating a semantic view of data centers, there are several works in the state of the art, many of which are ontology-based solutions where the logs are aligned to a specific available knowledge representation of the data center [4, 6, 12, 13, 15]. While in this work we also create a semantic representation of the data, we do not assume the existence of any target knowledge, but we let the semantic representation emerge from the human interaction: each time the subject matter expert adds semantic tags to characterize the logs, we collect and organize them in a growing taxonomy. The work by Mavlyutov et al. [14] is similar in this sense of not using any pre-existing target knowledge, but letting the semantic representation emerge from the data. The proposed Dependency-Driven Analytics (DDA) infers a compact dependency graph from the logs, which constitutes a high level view of the data that facilitates the exploration. Similarly, we generate high level, semantic views of the data center, but we do so by adding semantic tags to logs, with a human-in-the-loop paradigm rather than solely relying on parsing.

Last but not least, logs are not the only source of information available in data centers. A plethora of sources of data are available, including network data, configuration management, databases, data from monitoring devices and appliances, etc., all of which can be leveraged to generate a model of the data center [3, 5, 8]. While we use all this data, the major novelty of our work is that we combine the views derived from the data with the semantic tags iteratively added by the human exploring the data.

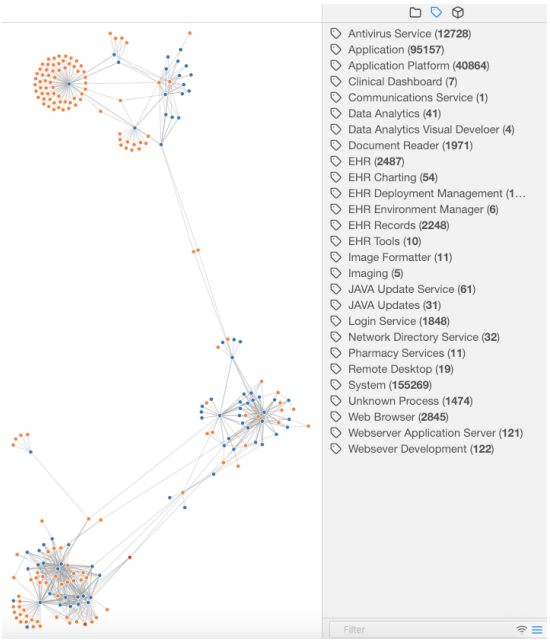


Figure 1: shows all nodes in the data center running a specific software (i.e. EHR) and all dependencies involving these nodes classified as being of type "EHR"

3 USE CASE: DATA CENTER EXPLORATION

During the demo we will showcase how we use the system to transition from raw data, to semantically enriched data, to full graph visualization and exploration of the data center.

3.1 Creating a Data Center Cartography

As a starting point we use all network information collected from the data center and build a color coded representation of all the nodes, as well as their incoming and outgoing connections. The created graph is (i) too big to be effectively visualized and (ii) does not contain crucial information about which processes are running on each node, to be able to design a migration plan. Therefore we collect information on running processes from all the nodes to try and characterize the nature of dependencies between each machine. The task is to effectively extract entities from each process log, where the entities of interest are the processes running on the machine. There are numerous hurdles to perform this task. First, the nature of the logs is single lines with a command which has been run on the machine and potentially a number of parameters. The format of the string is highly dependent on the operating system and on the specific command, but it is not a regular natural language sentence - therefore state of the art entity extraction tools fail. Regular expressions are helpful with identifying the meaningful parts of the process logs and matching them - when possible - to a list of candidate processes or applications. Nonetheless, not all processes are known a priori, therefore leaving us with numerous entities

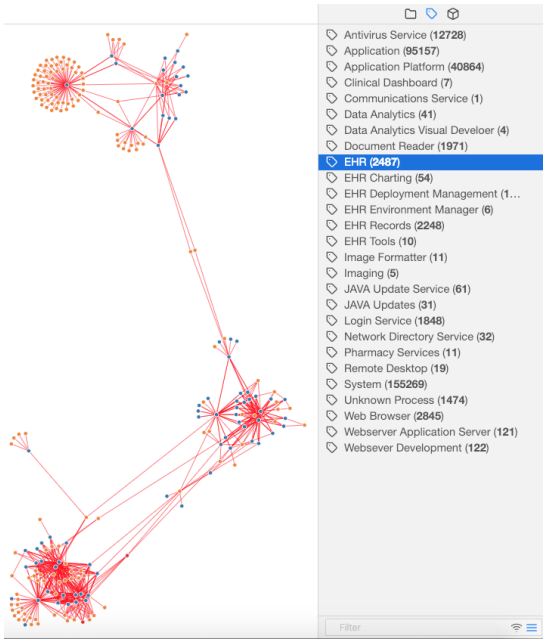


Figure 2: Example of selecting the tag "EHR" to highlight all dependencies that have been classified as being of type "EHR". All dependencies are highlighted in red.

which are nearly impossible to automatically resolve (the entities extracted from the logs can have obscure names, e.g. “xyz.exe”). We adapt our human-in-the-loop approach for corpus analysis [1] and perform the following approach: (i) we group logs that share the same entities and identify helpful clues in the parameters or in the directory paths (ii) we show these to the subject matter expert who can add a label if she recognizes which application is running. We then create a rule exploiting this knowledge and apply it to the whole graph. As the user keeps exploring and annotating the data, we dynamically add all the new tags as exploring dimensions, which can be immediately used to query and visualize the data. During the live demo we will show how we generate the graph from raw data and the effect of adding additional semantic tags - via the human-in-the-loop data enrichment approach - to the graph.

3.2 Exploring tagged dependencies

After the subject matter expert has interacted with the system, creating and enriching the cartography of the data center, the next step is to perform in-depth analysis. Specifically, when considering the task of planning the data center migration, one of the paramount questions to answer is to identify nodes that support the same (or dependent) processes, or as we refer to them in the following, “affinity groups". Finding a set of nodes that form an affinity group, followed by understanding the nature of the dependencies between nodes in the group is both valuable and challenging. An example of an affinity group may be a group of servers involved in providing and maintaining Electronic Health Records (EHR) from an EHR application. Figure 1 shows the initial step of surfacing an affinity group related to EHR. Each node is serving and/or consuming a service identified as EHR. Where nodes are connected by lines, we can say the nature of the dependencies between those nodes includes EHR (and likely many more classification as well). To emphasize this, Figure 2 shows the same cluster when the user selects EHR from the list on the right side of the screen, highlighting all dependencies involving EHR. This shows all tags involved in the current visualization, but given that the original query asked for a EHR affinity group, it follows that every dependency would be highlighted in red. To drill down further, a user can interactively select the other tags represented in the visualization to surface the dependencies where those tags are also involved. As Figure 3 shows, selecting "Network Directory Service" highlights a subset of the visualized dependencies. By visualizing the dependencies, tightly coupled groups are easily identified by the human eye. In the case of "Network Directory Service", one can immediately see there are 2 main nuclei, each with a set of non-overlapping dependent nodes. This sort of visualization also exposes insights that may not have been directly queried. The vast majority of business critical application groups will include a resiliency layer to provide redundancy, data backups, etc. However, the list of categories included in the EHR cluster does not include "Resiliency"or "Backup". To a user familiar with dependency grouping, this anomaly is striking. In fact, we learned on a call with the client that this particular cluster suffered an outage the week before, without any disaster recovery in place, effectively taking the whole business application offline.

4 CONCLUSIONS AND FUTURE WORK

Data exploration is a task that inherently requires a tight human interaction. The more complex the data and the scenarios, the more the need of a carefully designed methodology to support a human-in-the-loop paradigm. In this work we explored a particularly complex data exploration task, in the context of data centers, with the specific goal of supporting the subject matter expert to design a migration plan. Migrating a data center is a difficult and very critical task faced by many enterprises looking to increase flexibility, reduce costs, and enhance resiliency. The process requires a deep understanding of the dependencies between nodes, as well as the relations between larger clusters of nodes, either to each other and/or with other shared resources (eg. data stores, APIs, etc). Typically designing a migration plan can take 3-6 months, followed by 12-18 months for execution. Additionally, inevitable “missteps" along the way can cost substantial amounts of time and effort to resolve, along with the larger risk of business application outages. Our approach enables subject matter experts to obtain the necessary understanding of a data center in a faster and more accurate fashion. The tool has been used internally to help the subject matter experts in the design of migration plans, which has resulted in a positive reduction of design time and a significant reduction of mistakes/misunderstandings. The subject matter expert is provided with a clearer, more precise picture of the world they are working in, such that the job of planning and executing a smooth data center migration is far more realizable.

Future work for us focuses not only on reducing this “time to understanding”, but also assisting the practitioner in considering more complex parallel moves of functionality to the cloud, potentially reducing the migration time ever further. Given the nature

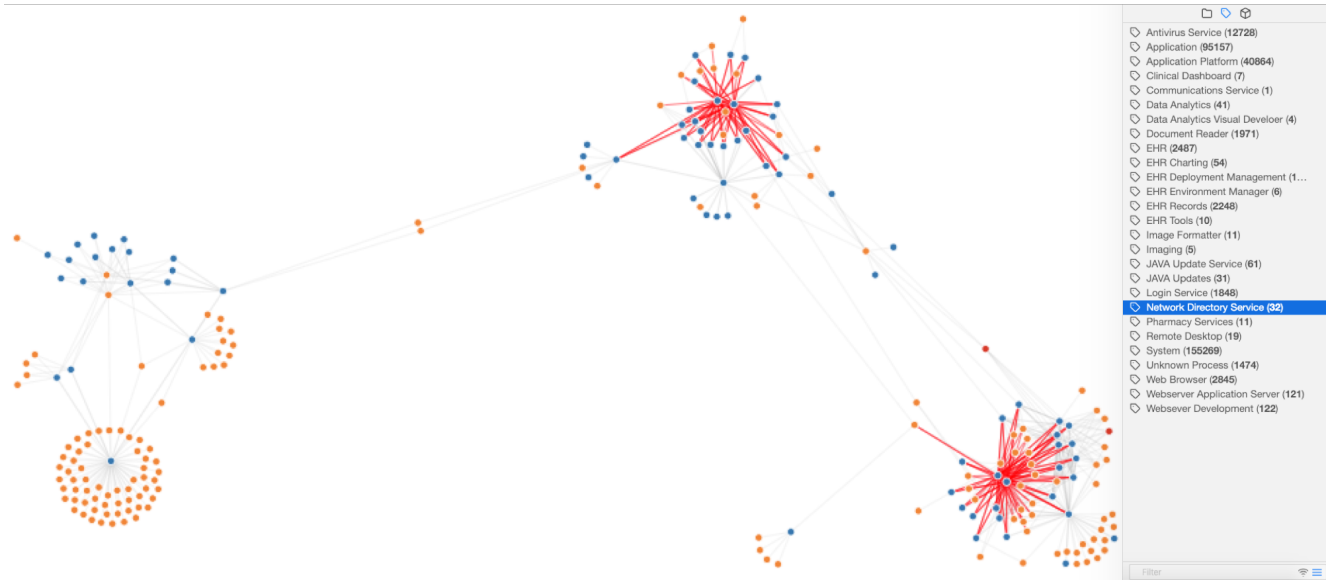


Figure 3: Example of selecting the tag "Network Directory Service" to highlight all dependencies that have been classified as being of type "Network Directory Service", a subset of dependencies are highlighted in red. The list of tags on the right shows only tags used in the visualization (i.e. the query results).

of our data model, there are opportunities for more advanced techniques (graph analysis and machine learning) to assist the human in annotating data. Automatic cluster suggestions using graph analysis is an area of active exploration. With the dependencies and annotations (tags) already in place, representing the data as a graph is a natural fit. Further analysis has the possibility of deepening our understanding of the relationships between nodes in such a way that rough identification of node clusters may become viable. This would involve starting with the nuclei and expanding outwards, only following paths of importance above a human-defined threshold until the analysis identifies a likely stopping point, thus firming the outer boundary of the cluster. While graph analysis looks at the nature of dependencies and relies heavily on tags, there are likely additional machine learning techniques that can be employed to enhance the tagging process, while keeping the human in the loop. This is an active area of exploration and will potentially speed up the tagging process significantly, but more importantly, broaden the set of entities with tags applied, thus providing a richer dataset to analyze.

REFERENCES

[1] Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. 2017. Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop. In *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017*, Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo (Eds.). ACM, 24:1–24:8. https://doi.org/10.1145/3148011.3148021

[2] Mahmoud Awad and Daniel A Menasc. 2015. Automatic Workload Characterization Using System Log Analysis. In *Computer Measurement Group Conf.*

[3] Zakaria Benzadri, Faiza Belala, and Chafia Bouanaka. 2013. Towards a formal model for cloud computing. In *International Conference on Service-Oriented Computing*. Springer, 381–393.

[4] David Bernstein, Santa Clara, Nipoma Court, and David Bernstein. 2010. Using Semantic Web Ontology for Intercloud Directories and Exchanges 2330 Central Expressway Using Semantic Web Ontology for Intercloud Directories and Exchanges 2330 Central Expressway. (2010).

[5] Robert H Bourdeau and Betty HC Cheng. 1995. A formal semantics for object model diagrams. *IEEE Transactions on Software Engineering* 21, 10 (1995), 799–821.

[6] Yu Deng, Ronnie Sarkar, Harigovind Ramasamy, Rafah Hosn, and Ruchi Mahindru. 2013. An Ontology-Based Framework for Model-Driven Analysis of Situations in Data Centers. (2013). https://doi.org/10.1109/SCC.2013.98

[7] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17, 3 (1996), 37.

[8] Tyrone Grandison, E Michael Maximilien, Sean Thorpe, and Alfredo Alba. 2010. Towards a formal definition of a computing cloud. In *Services (services-1), 2010 6th World Congress on*. IEEE, 191–192.

[9] David J Hand. 2007. Principles of data mining. *Drug safety* 30, 7 (2007), 621–622.

[10] Yongyao Jiang, Yun Li, Chaowei Yang, Edward M Armstrong, Thomas Huang, and David Moroni. 2016. Reconstructing Sessions from Data Discovery and Access Logs to Build a Semantic Knowledge Base for Improving Data Discovery. *ISPRS Int. J. Geo-Inf* 5, 5 (2016). https://doi.org/10.3390/ijgi5050054

[11] Mouad Lemoudden and Bouabid El Ouahidi. 2015. Managing Cloud-generated Logs Using Big Data Technologies. In *International Conference on Wireless Networks and Mobile Communications (WINCOM)*.

[12] Li Liao, Yuzhong Qu, and Hareton K N Leung. 2005. A Software Process Ontology and Its Application. In *ISWC2005 Workshop on Semantic Web Enabled Software Engineering*. 1–10.

[13] K Magoutis, M Devarakonda, N Joukov, and N G Vogl. 2008. Galapagos : Model-driven discovery of end-to-end application âĂŞ storage relationships in distributed systems. *IBM J. REs & DEV* 52, 4 (2008), 367–377.

[14] Ruslan Mavlyutov, Carlo Curino, Boris Asipov, and Philippe Cudre-mauroux. 2017. Dependency-Driven Analytics : a Compass for Uncharted Data Oceans. In *8th Biennial Conference on Innovative Data Systems Research (CIDR âĂŸ17)*.

[15] Lamia Youseff, Maria Butrico, and Dilma Da Silva. 2008. Toward a Unified Ontology of Cloud Computing. In *Grid Computing Environments Workshop, 2008. GCE '08.*

Uncovering Bias in Ad Feedback Data Analyses & Applications*

Marc Bron
Oracle

Ke Zhou
University of Nottingham

Andy Haines
Apple

Mounia Lalmas
Spotify

ABSTRACT

Electronic publishers and other web-companies are starting to collect user feedback on ads with the aim of using this signal to maintain the quality of ads shown on their sites. However, users are not randomly sampled to provide feedback on ads, but targeted. Furthermore some users who provide feedback may be prone to dislike ads more than the general user. This raises questions about the reliability of ad feedback as a signal for measuring ad quality and whether it can be used in ad ranking. In this paper we start by gaining insights to such signals by analyzing the feedback event logs attributed to users of a popular mobile news app. We then propose a model to reduce potential biases in ad feedback data. Finally, we conclude by comparing the effectiveness of reducing the bias in ad feedback data using existing ad ranking methods along with a new and novel approach we propose that takes revenue considerations into account.

ACM Reference Format:
Marc Bron, Ke Zhou, Andy Haines, and Mounia Lalmas. 2019. Uncovering Bias in Ad Feedback Data Analyses & Applications. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317304

1 INTRODUCTION

In today's society we often rely on free web services for many of our daily activities such as emailing, consuming digital content and social networking. One revenue model, prevalent among Internet companies providing such free services, revolves around sponsored posts or advertisements (ads for short). In this model, advertisers pay Internet companies to show ads to their users in a way that some will engage with the advertised message and thus deliver a return on investment to advertisers. Because of the self service nature and scale of these advertising platforms, a delicate balance needs to be struck between providing a rewarding investment to advertisers, whilst at the same time minimizing negative impact to users through caused by "bad" ads, or ads of poor quality that creep into the system.

In recent years, research has focused on using historical event logs or editorial judgements to develop implicit methods that ensure users are served relevant [2, 10, 11, 14, 18, 19, 26] and high quality ads [5, 30] and to define the impact and importance of both criteria [13, 15, 21]. These methods, however, are not perfect and some online companies have incorporated feedback mechanisms that allow users to express their opinion about ads explicitly. Ad feedback tools are used by companies like, Facebook, Twitter, Yahoo and YouTube. Such signals have the potential of improving the performance of existing quality models and deriving new ones.

Before applying feedback data to such scenarios, there are two important challenges to overcome. First, we must ask whether ad feedback as a signal is generalizable beyond individual users and representative of a service's user base. The fact that ads are targeted introduces the potential for the presence of bias in ad feedback signals. Figure 1 illustrate this. Moreover, users may not all respond equally to ads as some are more prone to provide feedback than others. Second, experimentation with ad quality signals, such as ad feedback, comes at a cost in terms of the short-term revenue of Internet companies [15]. Especially online services operating solely under the sponsored ad business model may be unwilling to experiment with ad quality signals to improve their users' experience when that may have a major revenue impact, e.g., a 50% reduction in ad impressions [15]. To allow online services keeping improving quality their is a need to be able to experiment with ad quality signals in a commercial setting while managing the potential revenue impact.

To address these challenges we start with an analysis of a large dataset of ad feedback data captured from the ad feedback mechanism of a large Internet media company. This mechanism allows users to provide feedback by hiding certain ads they are exposed to. We investigate to what extent the association between ads and ad feedback is affected by the fact that ads are targeted at users with particular demographics, interests, and behaviour. Furthermore, since user may differ in terms of their responses to ads we analyse to which extent user behaviour (e.g. clicks) affects feedback.

With this understanding we then start to answer the question as to where and how *bias* resides in feedback data as well as how feedback data may be incorporated in a commercial ad serving setting, while accounting for this bias. We develop a model that *corrects* for the bias in ad feedback data to produce ad quality scores that reflect the quality of ads in their truest sense. We then introduce a technique that provides explicit control over the revenue impact when incorporating ad quality scores in the ad ranking mechanism of an online ad auction. Finally, with all this in place we compare two models, that provide biased and bias corrected feedback estimates,

*This work was carried while all authors were at Yahoo Labs.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317304

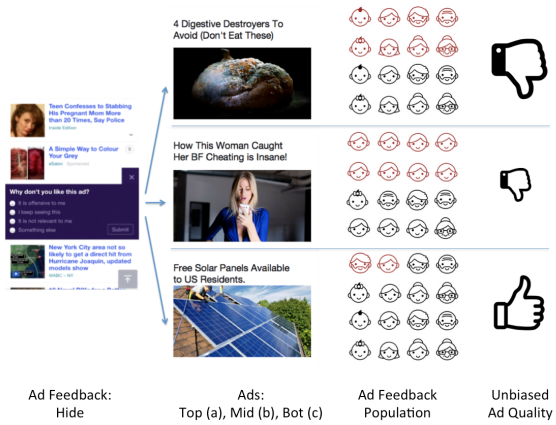


Figure 1: An anecdotal example of ad feedback vs. ad quality.

to investigate the value of correcting for bias in ad feedback data using various ad ranking strategies.

2 RELATED WORK AND MOTIVATION

Research in online advertising have mostly focused on predicting how an ad will perform e.g. [4, 6, 23, 24, 26]. The performance of an ad campaign is measured using a *score*, usually *click-through rate* (CTR), which is the number of times the ad was clicked out of the number of times it has been shown (number of ad impressions). Predicting CTR has been studied for many types of ad formats, e.g. [2, 4, 10, 11, 14, 18, 19, 22, 24, 26, 28, 29]. Additionally, in sponsored search, accounting for the relationship between the ad landing page, the query [6, 11, 19, 26] and the (dwell) time a user spent on the ad landing page [14] have been shown to be beneficial.

Each impression a user makes by visiting a publisher site the ad platform services is determined by an auction where the ad with the highest expected revenue wins and is thus shown to the user. Given a request for an ad to be served for that impression, ads are ranked accordingly through a real-time auction mechanism e.g. [3, 17], combining the bid - the amount of money the advertiser is willing to pay for its ad to be shown - and the score. This is formalised as $bid \times score$, where the latter usually corresponds to the predicted probability of the ad being clicked.

The publisher on the other hand may also want to restrict ads that could have detrimental effects to the long-term engagement of their users even if it means taking a potential hit in short term revenue. Whilst ads for the most part aim to provide users with useful information about products, at the same time they may annoy, distract or offend users.

In display advertising, [13] showed that "annoying ads have a real cost to users beyond mere annoyance", such that users develop negative attitudes toward host sites, causing reduced visits of shorter duration, fewer referrals, and overall *long-term* user disengagement. Ad features that related most to ad annoyance, include animation (e.g., motion), attentional impact (e.g., distracting), aesthetics (e.g., ugly), reputation (e.g., spam, fake), and logic (e.g., confusing, unfocused). In sponsored search, [15] showed the benefit of presenting users high quality ads (e.g. ad load time on mobile) on long-term business impact. Finally, regarding the negative effects of ads, [7] made important recommendations for both publishers and advertisers, with particular attention to the rise of ad blockers - a serious threat to publishers whose revenue largely relies on advertising.

In advertising, an important criteria of quality is how relevant the ads are. In sponsored search, approaches incorporating relevance in deciding which ads to return have been effective [11, 14]; in addition, [8] showed that it is better to not show any ads than to show non-relevant ones. Some works have focused on building models of ad quality, allowing to predict the probability for an ad to be bad, and to use the predicted probability in the estimation of the ad score, e.g. [5, 12, 21, 30]. For example, ads with very short time spent on their landing pages (high bounce rate) have been shown to be detrimental to long-term engagement, and when ranking ads with an ad quality component led to higher CTR [5, 21]. The message of the work discussed above is clear - returning ads of whatever definition of poor quality you chose results in negative consequences to long term revenue. It is for this reason that using explicit feedback mechanisms from users can help capture all these effects and once integrated directly into the ad ranking score allows ads to be ranked in terms of both *short-term* and *long-term* expected revenue.

Closest to our work is [30] who used ad feedback to train models to predict bad quality ads. Using the ratio of times an ad was marked offensive over the number of times it was shown to users (defined as "offensive rate") as an objective function they found the top predictive features to be trustworthiness, the product/service provided, the brand and layout, aligning with those reported in [7, 13]. However, in their work they did not take into account the fact that different ads are targeted at different users. Nor did they consider that users differ in their tendency to provide feedback. In other words, users may dislike ads but not indicate this through a feedback option whereas others may always give feedback, however minor the complaint. These potential sources of bias imply that reducing feedback may not equate to improving ad quality for all users.

Any ranking model that utilises such an explicit signal needs to take care when penalizing ads based on a vocal minority. This raises the following two challenges for ad feedback to be used as a quality signal in online advertising applications:

(1) Not all users provide feedback. Can the feedback given be considered reflective of the entire population?
(2) Given the penalization nature of feedback signals in ranking there is naturally a cost attributed in the form of lost short term revenue. Given that publishers will be unwilling to sacrifice such revenue loses, is there a way of finding a tunable compromise between this cost and the quality improvements?

3 AD FEEDBACK DATA

In this paper we use data collected from the feedback mechanism of a popular mobile news app from a major online company.[1]

To illustrate the setup, consider the three anecdotal ads and corresponding feedbacks as shown in Figure 1. Here, users can

[1]The news app was Yahoo news. However the work presented here is relevant to any online sites where ads are served, and an ad feedback mechanism is in place.

Table 1: Variables and levels for each variable used to partition users. The level *unknown* is used when missing data is present.

Variable	Levels	Dim
	User demographics level partition	
age	below 24; 25-29; 30-34; 35-44; 45-54; 55-59; 60-64; 65+; unknown	9
gender	male; female; unknown	3
location	50 US states	50
interests	business; telecommunications; health; travel; automotive; entertainment; sports; finance; technology; retail; unknown	11
	User behaviour level partition	
impressions	indication of exposure to ads	25
ad clicks	indication of interaction with ads	25
article clicks	indication of interaction with articles	25

choose to hide ads they are negatively impacted by simply clicking on the small 'x' appearing in the top right-hand corner of the ad. In the feedback population column we illustrate the types of users who saw a given ad and colour them red when they gave feedback versus black for those who did not. Since the top two ads (a and b) are hidden by more users per impression, they are indicative of a poorer ad quality than the bottom ad (c). However, say for argument's sake that teenage males overall provide more feedback than the rest of the population. For ad (b), since this feedback is attributed only to teenage males, the quality of ad (b) should be higher than ad (a) as the negative feedback attributed to (a) is from more diverse segments of the population.

We extracted a random sample of ad feedback data collected over a two week period from from the US version of the news mobile app. Our sample contains around 40 million distinct users and 200,000 distinct ads. Since feedback rate is a business sensitive metric, we characterize the ad feedback rate relative to ad CTR. The hide rate is 84.0% smaller than CTR, where hide rate is defined as follows:

$$HR = \frac{\#(\text{hides})}{\#(\text{impressions})}$$

where $\#(hides)$ is the number of times users clicked to hide an ad and $\#(impressions)$ is the number of ad impressions.

One may argue that CTR is already an indicator of ad quality. However, feedback is generally a signal of bad quality while the absence of a click does not necessarily indicate a low quality ad. In addition, high CTR may not necessarily mean high quality. As discussed in [30], many ads labelled as offensive could be seen as "provocative", hence attracting clicks. Through hides the user provides an explicit negative signal. A first contribution of this work is to provide insight into whether this signal is representative across users (whether only particular users provide hides), ads (whether only bad ads receive hides) or a combination of these.

The variables (features) and their levels (feature values) used to characterize users are described in Table 1. We distinguish between user demographics and user behaviour based variables. User demographic variables age and gender, are based on users' self declared information (when signing up to the news app). User interests are inferred based on articles clicked by users in the last month using a proprietary in house algorithm. Location is based on the most recent IP address observed for a user.

We categorize user behaviour in terms of the number of ad impressions, ad clicks, and news item (article) clicks. Impressions are a proxy for engagement with the news app, whilst ad clicks and article clicks are proxies for engagement with specific ads or articles, e.g., a user may skim headlines but never read any articles. We bin the counts for each variable on a logarithmic scale into 25 categories.

Regarding the features to characterize ads, we were inspired by [30] and selected three types of ad features specific to the ad copy.

- *Text-based* features are derived from the title and description of an ad and include: (a) *spam*, the extent in which the text has spam keywords; (b) *readability*, the readability level according to the Gunning fog index (primary school, high school, etc); (c) *adult*: whether the product advertised is adult-related e.g. dating sites.
- *Image-based* features are derived from any image used in an ad and include: (d) *image text detector*: whether the ad image contains text; (e) *flesh*: likelihood that an image contains adult content (e.g. too much skin).
- *Advertiser* features are derived from the advertiser that placed the ad and include: (f) *brand*: the pagerank score of the top level domain of the ad landing page, reflecting its popularity as a brand.

4 AN ANALYSIS OF AD FEEDBACK

We start by exploring the relationship between user variables and hide rate. To understand which variables may be used to characterize users that hide ads, we study the difference of ad hide behaviour for each user variable in turn. Within each variable e.g. user interests, we define a *cohort* per level, e.g., Travel, aggregate the feedback across each cohort and compare the cohorts. If users within a cohort are targeted with particularly low quality ads compared to other cohorts then we would expect the hide rate to be higher than average for that level of a user variable.

We visualize how the hide rate of various user cohorts $HR(u)$ differ relative to the mean ad hide rate of all user cohorts U. This is defined as:

$$HR_{var}(u) = \frac{HR(u) - mean(HR(U))}{mean(HR(U))}$$

By visualizing the difference from the mean across levels of each variable we gain insight in whether such a variable may be involved in targeting, e.g., users interested in technology may be more likely to provide ad feedback than users interested in travel, or represent a group of users sensitive to ads.

User Demographics. We first plot the relative difference from the mean HR_{var} according to various US states in Figure 2. We observe that users from different states exhibit differences in hiding ads. This is likely due to targeting as location is a popular targeting criterion. However, demographic distributions differ per state and a generally younger or older state level population may affect hide rate as well.

We next plot HR_{var} across different user age and gender groups in Figure 3a. We observe that female users are less likely to hide

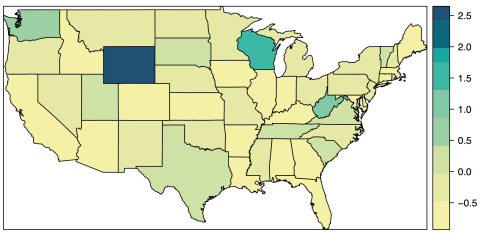
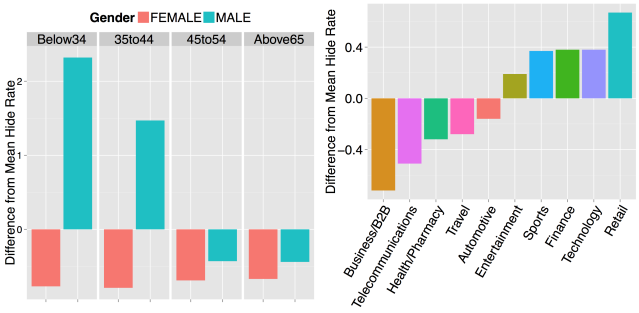


Figure 2: Ad hides across US states



(a) Ad hides across age and gender

(b) Ad hides and user interests

Figure 3: Relation between hides and user variables, i.e., age by gender, and interests.

ads than male users. With respect to age, young male users tend to hide ads more often, whereas there are no differences across the age groups for female users. Finally, young male users and young female users below the age of 44 are, respectively, most and least likely to hide ads.

User Interests. We plot the HR_{var} across different user interest groups in Figure 3b. We find that users interested in "Retail" and "Technology" are most likely to result in hidden ads. In contrast, users interested in "Business/B2B" and "Telecommunication" are the least likely to provide ad feedback. As with the demographics variables, interest variables are popular targeting criteria and advertisers may target based on one or all of these. We find that feedback variations across the levels of these variables make them suitable candidates to identify bias due to targeting.

User Engagement Patterns. We are also interested in whether users with different engagement levels with respect to the host content platform respond differently to the ads shown in terms of feedback. We categorize users into four quantiles according to a given engagement metric, where "Q4" represents the highest quantile (the most engaged users) and "Q1" denotes the lowest quantile (the least engaged users).

We use three metrics of user engagement: users' ad impressions, ad clicks and content item (e.g., article) clicks. The first metric is a proxy of how engaged a user is with the news app, as the more engaged, the more the user will have seen ads. The second metric characterizes the extent to which a user engages with an ad, i.e. clicking on it. The final metric is the extent to which the user engages with the content items, i.e. reading news articles.

Table 2: User engagement levels and ad feedback

Quantile	Ad Impressions	Ad Clicks	Item Clicks
Q1	-0.38	-0.73	-0.85
Q2	1.89	0.85	1.14
Q3	-0.65	-0.96	0.26
Q4	-0.87	0.84	-0.55

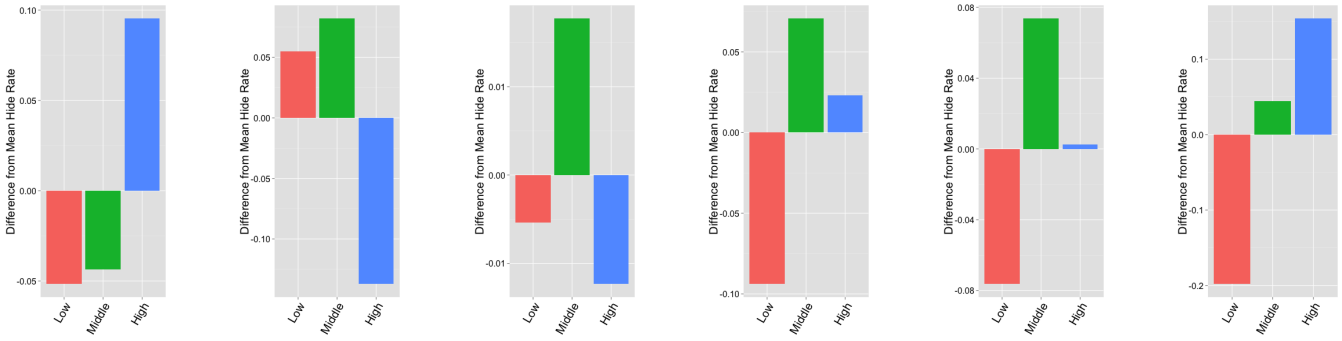
Table 2 shows how each user behavior metric, at various engagement levels, affects ad hide rates compared to the mean across all users. We observe first that users who are the least engaged with the service and with the ads, are less likely to hide ads (quantile "Q1"). We also see that the second level of engaged users, with respect to all metrics, are those that are more likely to provide ad feedback, i.e. they are more likely to hide ads ("Q2"). Finally, users that are highly engaged with the service tend to hide less ads, but those that interact with ads do provide more ad feedback ("Q4").

The above behavioural variables are not available as targeting criteria. Nonetheless we still observe considerable variations across the levels. For example, those users that sometimes saw ads (Q2 of *ad impressions* in Table 2) are almost as twice as more likely to hide the ads. One explanation is the presence of a response bias for users with different engagement levels, e.g., a user highly engaged with a content platform may be less sensitive to ads shown there than less frequent users affecting their propensity to provide feedback.

Ad quality. High and low values of ad features aimed to characterize the quality of ads should be associated with different levels of ad feedback in order to be useful. Here, we explore the relationship between different levels of ad features and hide rate. We derive ad features from the ad copy (ad text, image and advertiser) and investigate how variants of each affect hide rate. Similarly to our user cohort analysis, we study the difference in hide rates with respect to each ad feature relative to the mean hide rate.

We use the ad features described in Section 3. For each ad feature, we separate all the ads into three equally sized bins ("Low", "Middle" and "High") and visualize how the ads in each bin differ from the mean hide rate of all ads. The results are presented in Figure 4. For text-based features, the more likely an ad has its text identified as *"spam"*, the more likely the ad is hidden (Figure 4a). By contrast, the *easier* the ad is to be *read* (e.g. written at a level associated with primary school students), the more likely it is being hidden (Figure 4b). Finally, ads with the most and least *Text adultness* are less likely to be hidden (Figure 4c), while ads with a moderate level of adult content are most often hidden. In terms of image-based features, the more likely an ad contains *"text"* (Figure 4d) or *"flesh"* (Figure 4e) in its image, the more likely it will be hidden.

Summary. We find that both the user variables and ad features are associated with different levels of ad feedback. Apart from the ad quality, some types of users that provide feedback may be more sensitive to ads and have a higher tendency to provide feedback. In addition, since ads are targeted, the ad feedback received maybe from a group of users unrepresentative of the general population. In the next section, we model this bias considering both users'



(a) Text Spamness (b) Text Readability (c) Text Adultness (d) Image Text (e) Image Flesh (f) Advertiser Brand

Figure 4: Understanding the relationship between ad features and hide rate.

characteristics (those popular ad targeting criteria, such as age and interests) and their behaviour (e.g., their engagement levels).

5 MODELLING BIAS

The previous analysis provides insights into the distribution of ad feedback across individual user and ad characteristics as well as the presence of bias due to user ad sensitivity and targeting in the ad feedback data. An ad quality model based on such biased data will consistently over- (or under)-estimate the quality of ads. To account for such bias we first, in this section, develop a model able to determine the proportion of bias present in the feedback on ads. Then, in Section 6, we use this model to develop an ad quality model based on ad feedback data corrected for bias.

5.1 Approach

Ideally one would determine the quality of ads based on ad feedback through ad impressions randomized across users. After sufficient impressions, an ad feedback rate would be an unbiased estimate of its true quality. However, since ads may be targeted at users based on specific characteristics or shown to users more prone to give feedback such data is not available. Hence, we have a sample selection bias problem where ads are shown to a selected sample that nay not be representative of the population.

To correct for this selection bias we do not aim to develop the best *predictive* model for ad feedback, nor do we need to experiment with various features and machine learning models. Although such a model can accurately predict users' feedback it implicitly models the bias present in the data. Instead we take an econometric approach [27] in which we develop a *descriptive* model [25] of user feedback behaviour that only includes variables able to explain the sources of selection bias, i.e., ad sensitivity and targeting. We *will* use this model to account for bias in an *ad quality model* capturing the relation between ad characteristics and ad feedback in Section 6.

To correct for selection bias we use the following procedure. First, consider a simplistic logistic regression based ad-user model with one ad feature a, one user selection bias feature u, associated weights w_a, w_u respectively, the intercept term w_0, the logit link function $f(.)$, and the model error ϵ:

$$f(\hat{p}) = w_0 + w_a \cdot a + w_u \cdot u + \epsilon$$

Further, we consider an ad only model using only one ad feature:

$$f(\hat{p}) = \hat{w}_0 + \hat{w}_a \cdot a + \epsilon$$

If both models are fit to the feedback data with the selection bias present then the bias in the coefficient of the ad model can be characterized as follows:

$$\hat{w}_a = w_a + \rho w_u,$$

where ρ is the correlation between ad feature a and user selection bias feature u. That is, the bias in the ad only model is the true user bias proportional to the correlation between the user and the ad feature. In contrast, in the ad-user model the user selection bias is modelled explicitly and the true effect of the ad feature is captured by w_a, whereas the user selection bias is captured by w_u. Using the w_a and intercept terms from the explicitly biased model, we obtain:

$$f(p) = w_0 + w_a \cdot a + \epsilon$$

This is the feedback rate predicted purely based on the ad characteristics and with weights that are no longer conflated with biased user feedback. Next we develop the user selection bias model and investigate its properties.

5.2 The user selection bias model

We start with developing a model to identify user selection bias term ($w_u \cdot u$) in the previous equation and then analyze how combinations of characteristics relate to the presence of selection bias in ad feedback. We consider the user characteristics from Table 1 for inclusion in our models and employ a forward model selection based strategy to determine the composition of the final user bias model. As criteria for model fit we use the deviance statistic [16]:

$$D = 2 \sum y_a \log \frac{y_a}{\hat{y}_a} + (N_a - y_a) \log \frac{N_a - y_a}{N_a - \hat{y}_a}$$

where y_a is the observed number of hides and $\hat{y}_a$ is the predicted number of hides for group a.

Table 3 shows the deviance statistics for the models of interest ranging from the null to the best fitting model. The first column lists the model name, the second uses notation from [9] to specify the systemic structure of the model, the third column lists the deviance statistic, the fourth column the difference between the null model

Table 3: Deviance statistics for models of interest ranging from the null to the best fitting model for ad hide data of a news app.

model ($\mathcal{M}$)	notation	deviance	$\delta_{\emptyset-\mathcal{M}}$%	param
Null	$\emptyset$	1438444	-	1
age	a	978274	32	10
gender	g	1413902	2	4
state	s	1169876	19	51
interests	i	788621	45	12
adclk	ac	992030	31	26
itmclk	ic	959711	33	26
adimp	ai	1184212	18	26
additive	a+g+s+i+ac+ic+ai	323621	78	143
interactions	a+s+i+ac+ic+s:i	304426	79	607

and the current model, and finally the fifth column lists the number of parameters used in the model.

We observe that the age variable reduces deviance by 32% (column $\delta_{\emptyset-\mathcal{M}}$) using ten parameters. By contrast, gender reduces deviance by 2% using four parameters and the state variable reduces deviance by 19% using fifty-one parameters. These suggest that different age levels are more indicative of different hide rates then gender or state. However, we find that the interest variable has the best model fit and reduces deviance by 45% using twelve parameters suggesting that interest variables are a popular targeting criteria.

The behavior based variables ad click (adclk) and item click (itmclk) are similar in reducing deviance by 31% and 33% using twenty-six parameters, whereas ad impressions (adimp) reduces deviance by 18%. Moving to the additive model, i.e., combining all variables but not their interactions, reduces deviance by 78% on 143 parameters. However, the best fitting model including two-way interactions achieves a 79% reduction in deviance on 607 parameters. This model does not include gender and ad impressions, but does include an interaction term for state and interests.

Each of the reductions in deviance by adding variables to the models leads to significant improvements as determined using a χ^2 goodness of fit test - i.e., the variables included in the model are able to explain significant structure in the data. There is a meaningful amount of feedback that can be explained by targeting criteria, specifically, age as well as combinations of different levels of state and interest variables regardless of the ad characteristics. These suggest that there is a selection bias due to targeting present in our feedback data. We further observe that click behaviour variables are able to explain additional feedback, suggesting the presence of selection bias due to user ad sensitivity. Next, we describe the effect of specific values of these variables on selection bias.

5.3 User selection bias model inspection

The user selection bias model includes variables that can explain two sources of selection bias, i.e., targeting and ad sensitivity. We now inspect the model to understand how individual variable level weights positively or negatively affect the selection bias present in ad feedback - the *net effects*. We discuss the effects in terms of the odds ratio (odds for short) as it facilitates interpretation of the effect of each variable level on ad hides. We derive the odds ratio

Table 4: Net effects of variables in the best model. Significant effects are denoted by $^{\triangle}$($^{\blacktriangle}$) at the $\alpha < 0.05$ ($\alpha < 0.01$) level.

variable	level	net effect	variable	level	net effect
	Intercept	-21.1	adclk	24	-15.7
state	Maine	-0.782$^{\blacktriangle}$		11	2.01
	Georgia	1.249$^{\blacktriangle}$	itmclk	24	-0.553
interest	Entertain.	-0.104$^{\triangle}$		1	-1.39
	Retail	0.072	state:interest	OR:Auto	0.898$^{\blacktriangle}$
	Finance	1.35 $^{\blacktriangle}$		MT:Auto	0.992$^{\blacktriangle}$
	Auto	-0.206		NM:Telecom	1.124$^{\triangle}$
	Telecom	-0.759$^{\blacktriangle}$		SD:Retail	1.637$^{\blacktriangle}$
				WY:Auto	1.760$^{\blacktriangle}$

from a net effect β using:

$$\frac{p}{1-p} = e^{\beta}$$

When relevant, to determine the probability of hiding ads, we use the inverse logit:

$$p = \frac{e^{\beta}}{1+e^{\beta}}$$

Table 4 shows the effects of the levels of the variables as produced by the best fitting model. The first column lists the *variables* included in the model, the second column contains the *levels* of each variable, and the third column presents the *effects* of each level.

The intercept is negative and the odds ratio of hiding ads when controlling for all other factors is small ($6.86 \cdot 10^{-10}$). It reflects the hide rate of a user with no targeting characteristics and having the lowest engagement level in terms of ad click, item click, and ad impression behaviour. We will refer to this user as our reference user on the news app. This user's hide rate is orders of magnitude smaller than the general feedback rate (see Section 3) and implies that the reference user is unlikely to hide ads. The characteristics of our reference user are based on the levels that are used as reference cell for each variable, i.e., a user with no targeting characteristics and having the lowest engagement level in terms of ad click, item click, and ad impression behaviour. We discuss the variables related to targeting and user ad sensitivity each in turn.

Targeting effects. Table 4 shows the two states with the largest negative and positive effects. We observe that the net effect of changing our reference user state to Maine reduces the odds of observing a hide by 64% ($1 - e^{-0.782}$), whereas changing to Georgia increases the odds by 249% ($e^{1.249}$). This is in line with our observations in Section 4, Figure 2. In terms of the interest levels we find that an interest in Sports and Finance increases the odds of observing a hide by 150% and 284%, respectively. In contrast, interest in the Telecommunications category reduces the ad hide odds by 63%. These observations are again in line with our findings in Section 4, Figure 3b. We further discover that there are interaction effects; Table 4 shows the five levels with the highest positive net effect. Three of the five highest effects are combined forces between an interest in the Automotive category and some of the states, i.e., increasing the ad hide odds by 145% in Oregon, 151% in Montana, and 481% in Wyoming, respectively.

Behavioural effects. The behavioural variables capture the extent in which users interact with ads and items (e.g., articles). The interaction data for each user is binned on a logarithmic scale in 25 categories. Table 4 shows the 2 levels with the highest negative and positive effects. Moderate ad clickers are more likely to hide an ad than users with no ad click activity or very high ad click activity.

Further, there is an inverse linear relation between the number of ad hides and the 25 log scale levels of the itmclk variable (Pearson $r = -0.99, p < 0.001$). This implies that, unlike for ad clicks, users clicking on more articles are less likely to hide ads, whereas users clicking on few articles are more likely to hide ads. A user having an itmclk level of 24 reduces the odds of an ad hide by 42% and a user with an itmclk level of 1 increases the ad hide odds by 300%.

Discussion. Inspection of the user selection bias model shows that not all users hide ads equally as user characteristics alone are sufficient to explain a significant part of the feedback observed in our data without considering features indicating the quality of ads. Ad feedback varies depending on particular user targeting characteristics, which demonstrates that ads from certain ad campaigns with targeted, for example, towards users in specific locations with particular interests received high levels of feedback. Using feedback only from such particular groups of users is not appropriate to train a general ad quality model that determines whether ads are of high or low quality for the general population. Feedback also varies depending on the engagement level of users of the news app; for example, users with high engagement levels with articles were less likely to provide feedback than users with low engagement levels. It seems reasonable to trust the feedback received from users over various levels of engagement more than feedback only provided by a group of users with low engagement levels.

Next, we incorporate the user selection bias model into an ad quality model that uses the user selection bias as well as ad features to determine the hide rate of ads.

6 CORRECTING BIAS IN AD FEEDBACK

To correct for selection bias in the feedback used by our ad quality model we explicitly model the user selection bias in addition to the ad features as follows:

$$f(\hat{p}) = w_0 + w_a \cdot a + I(w_u \cdot u) + \epsilon$$

We use an indicator function to binarize the user selection bias term based on a threshold parameter $I : f(.) > t$. This parameter indicates whether feedback received on an ad is likely due to selection bias.

We consider four ad characteristics as variables in our model, namely, pagerank of domain (pr), adult score (a), readability description (r), and spam score (s). Given each ad feature distribution we separate them into equally sized bins representing the different levels. We use 100 levels for pagerank of domain, i.e., L1 to L100, 5 levels for the readability score and 10 levels for the remaining features. We do not include the ad image features as their distributions are more skewed, i.e. only a small number of ads contain text or flesh. Further, we set the threshold for the user selection bias model to 0.5 to indicate when part of the feedback received is likely to be due to selection bias and encode this as level L1 when $f(w_u \cdot u) > 0.5$ and L0 otherwise. Given the above variables we again use forward stepwise model selection based on AIC and allow inclusion of the main effects and all pair-wise interactions in the candidate models.

Table 5: Selection of the effects of the ad and user variables in the best fitting model for the ad hide data. In formula notation the model is specified as follows: u+p+a+r+s+u:p+u:a+p:a+p:r+p:s+a:r+a:s. Significant effects are denoted by $^{\triangle}$($^{\blacktriangle}$) at the $\alpha < 0.05$ ($\alpha < 0.01$) level.

variable	level	effect	variable	level	effect
	intercept	-9.203 ▲	user type:adult score	L1:L10	-1.144 ▲
user type (u)	L1	3.012 ▲		L1:L8	-0.632 ▲
readability (r)	L1	0.082 △		L1:L6	-0.586 ▲
	L4	0.204 ▲		L1:L2	-0.217 ▲
adult score (a)	L9	-0.187		L1:L1	-0.158 ▲
	L8	0.967 ▲	user type:pagerank	L1:L79	-2.498 ▲
spam score (s)	L7	-0.194 ▲		L1:L89	-0.401 ▲
	L9	2.058 ▲	pagerank:adult score	L79:L8	-3.478 ▲
pagerank (p)	L89	-0.993 ▲	pagerank:spam score	L79:L7	0.7758 ▲
	L79	3.270 ▲			

The effects of the variable levels of the resulting model are presented in Table 5. The model includes the main effects of user bias (u), page rank of domain (p), adult score (a), readability (r), and spam score (s). In terms of interaction effects, the model includes all pairwise interactions between page rank and the other variables (u:p, p:r, p:s, p:a,) and all interactions between adult score and the other variables (u:a, a:r, a:s). We observe that the hide rate derived from the intercept is 0.01%, which is several orders of magnitude higher than the overall feedback rate, see Section 3. This indicates that ads with low pagerank and readability levels as well as low adult and spam levels are considered of low quality by users when correcting for selection bias. Adding the effect of the user selection bias variable (3.012) results in a further 20 fold increase in the odds of observing a hide, suggesting that users specifically targeted by ads with such features and users with a high disposition to give feedback also consider such ads of low quality.

With respect to the other main effects, we find that the effects have mixed directions across the levels of each variable. The readability levels with significant effects all result in an increase in the odds of observing a hide; however, lower levels do less so than higher levels. The highest and lowest significant effects for the adult score are both for high levels of adultness. A similar observation holds for the spam score and pagerank. These observations reflect the presence of interaction effects between levels of ad features, i.e., single features do not characterize the quality of an ad.

In terms of interaction effects we observe that ads with some features are less likely to receive biased feedback than others. For example, an ad with adult level L8 increases the odds of receiving a hide by 163%; however, if the ad is impressed to a user of level L1 then the odds are reduced. Similarly, an ad with a high pagerank level, i.e., L79 increases the odd of receiving a hide by a factor of 36; but impressed to a user of level L1, those odds reduce by 92%. Ads with such features are likely to receive feedback from a general population, but unlikely to receive feedback from a specific segment of users. This may be indicative of ads with products or content prohibited in certain countries, demographics, etc.

The effects for the ad features in the bias corrected ad model determine the inherent quality of an ad as determined by the hide rate estimated for the general population and not a specific segment of users. By adding all effects based only on the characteristics of each ad — excluding the user selection bias factors and interactions — we obtain the unbiased hide rate of an ad that controls for the type of users that ad has been shown, i.e., $f(p) = w_0 + w_a \cdot a + \epsilon$, from Section 5. In the next section we show the utility of correcting for user selection bias in ad feedback in ad ranking.

7 APPLYING FEEDBACK IN AD RANKING

We first introduce the general ad ranking approach for online ad auctions as well as how ad quality scores are generally incorporated in such rankings. Then we describe our method that allows the use of ad quality scores in ad ranking under a fixed revenue constraint. With these in place, we compare the value that biased and bias corrected feedback estimates provide in terms of discounting or filtering bad quality ads with particular attention to how this impacts revenue.

7.1 Ad auctions and feedback data

Online advertisement impressions are sold using real-time auctions. In such auctions each time a user loads a webpage one or more ad slots become available for sale. Advertisers state their bids for an impression of one of these slots and ads are ranked according to a function of their bids. The ads with the highest ranks are displayed and if a user clicks on the ad the advertiser is charged an amount based on the bid of the ad directly below it in the ranking [3]. As it directly influences prices charged to advertisers and therefore the revenue of the host site, the ranking function is central to the auction.

Ads are generally ranked by expected cost per impression (eCPI), which is determined based on the probability that the ad (a) is clicked given a user impression (u):

$$\text{eCPI} = \text{bid}_a \cdot P(C_a = 1|U = u)$$

The immediate revenue is maximized by selecting the ad a from a set of ads A eligible for an impression during a real time auction with highest eCPI:

$$\arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|U = u) \tag{1}$$

where the probability of a click on ad a given user u may be estimated based on historical click data and preferences of similar users [1].

Some Internet companies have started to incorporate quality scores into their ad ranking mechanisms. We define the probability that an ad will provide a good quality experience as: $P(Q_a = 1|U = u) = 1 - p$, where p is an estimate that the user u will hide the ad. One common method to incorporate quality into selecting which ad to show to users is to discount the expected CPI by the probability that the ad will deliver a quality experience [20]. In the case of ad feedback we can formulate discounted eCPI as:

$$eCPI_q = \text{bid}_a \cdot P(C_a = 1|Q_a = 1) \cdot P(Q_a = 1|U = u)$$

and select the ad that maximizes revenue:

$$\arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|Q_a = 1) \cdot P(Q_a = 1|U = u) \tag{2}$$

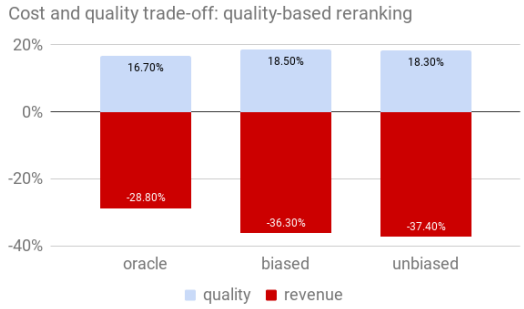


Figure 5: The % change in hides prevented (grey) and revenue delivered (in red) for three ways of estimating $P(Qa = 1|U = u)$: (i) *oracle*, using emperical log based estimates, (ii) *biased*, which uses predicted estimates based on the biased model $f(\hat{p})$ from Section 5; and (iii) *unbiased*, which uses predicted estimates based on corrected model from Section 6.

Results. We use two weeks of ad interaction and revenue data from the logs of the Internet company for which we study the ad feedback data to demonstrate the impact of incorporating feedback data (quality) in ad ranking. We compare the revenue delivered and negative feedback (hides) prevented over a two week period for three ad feedback based quality ranking methods (implemented using equation 2) with the performance of the current eCPI based model (i.e., equation 1) running in production of the Internet Company.

Figure 5 shows the % change in quality improvement (hides prevented in grey) and revenue delivered (in red) for three ways of estimating $P(Qa = 1|U = u)$: (i) *oracle*, which uses the emperical estimate based on the logs in hindsight, (ii) *biased*, which uses estimates based on the biased model $f(\hat{p})$ from Section 5; and (iii) *unbiased*, which uses estimates based on corrected model from Section 6. We observe that the general approach to incorporating quality scores into the ad ranking function does not deliver a favorable trade-off in terms of quality improvement and revenue investment. The quality ranking approach using oracle estimates of the quality score (hide rate) results in a 28.8% reduction in revenue, while reducing the number of hides by 16.7%. The biased and unbiased estimate based approaches achieve similar performance and are worse reaching a 36% and 37% revenue investment respectively for a 18% reduction in hides.

Next, we explore alternative ad ranking approaches that take revenue into account.

7.2 The revenue-feedback trade-off

When a publisher maximizes expected CPI that is discounted by the quality of ads it no longer maximizes revenue, unless the quality is proportional to eCPI or uniformly distributed. Specifically, given a set of ad candidates $A = \{a_1, ..., a_n\}$ for an impression we define the maximal expected revenue as:

$$\text{eCPI}_{max}(A) = \text{bid}_{a_i} \cdot P(C_{a_i} = 1|U = u)$$

where $a_i = \arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|U = u)$.
Further, we define the quality based expected revenue as:

$$\text{eCPI}_q(A) = \text{bid}_{a_j} \cdot P(C_{a_j} = 1|U = u)$$

where

$$a_j = \arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|Q_a = 1) \cdot P(Q_a = 1|U = u)$$

Then the loss in eCPI is equal to

$$\text{eCPI}_{loss}(A) = \text{eCPI}_{max}(A) - \text{eCPI}_q(A)$$

The gain of showing quality ads depends on an increase in impressions and clicks compared to not showing quality ads over a certain period of time. However, any increase in impressions and clicks, and therefore any benefit of showing quality ads, can only be observed in hindsight. We define the set of impressions that would have been observed while maximising for expected CPI during window (t, t_k) as $I_{(t,t_k)}$ and the impressions potentially observed while discounting for quality as $I_{q(t,t_k)}$, where we assume $I_{(t,t_k)} \subseteq I_{q(t,t_k)}$, i.e., quality ads indeed lead to more engaged users. Publishers then wish to find a time t_k where:

$$\sum_{j \in I_{q(t,t_k)}} \text{eCPI}_q(A_j) - \sum_{i \in I_{(t,t_k)}} \text{eCPI}_{max}(A_i) \geq 0$$

By splitting the impressions in those that would have been observed under the regular ad serving model $I_{(t,t_k)}$ and potential new impressions $I_q \setminus I$ we obtain:

$$\sum_{j \in I_q \setminus I} \text{eCPI}_q(A_j) + \sum_{i \in I_{(t,t_k)}} \text{eCPI}_q(A_i) - \sum_{i \in I_{(t,t_k)}} \text{eCPI}_{max}(A_i) \geq 0$$

Rewriting we find that publishers face the dilemma of whether to risk a loss in revenue

$$\sum_{i \in I_{(t,t_k)}} \text{eCPI}_{loss}(A_i)$$

in order to find if there is a time t_k where:

$$\sum_{j \in I_q \setminus I} \text{eCPI}_q(A_j) \geq \sum_{i \in I_{(t,t_k)}} \text{eCPI}_{loss}(A_i) \tag{3}$$

while having no control over the revenue loss

$$\sum_{i \in I_{(t,t_k)}} \text{eCPI}_{loss}(A_i).$$

7.3 Bounding short-term revenue loss

Whether the left hand side of Equation 3 will become greater than the right hand side depends on the relation between ad quality and user engagement as well as the magnitude of the bids associated with quality ads. To allow publishers to investigate whether quality ads could yield more revenue in a particular market place while being in control of the amount of revenue lost, we propose the following:

1) Select all the ads A that are eligible to be shown to the user;
2) Compute the maximum expected revenue given A: $\text{eCPI}_{max}(A)$;
3) Find the ad with maximum user experience such that the loss in revenue is smaller than a threshold τ:

$$\begin{aligned} a_c &= \arg\max_{a \in A} \ P(Q_a = 1|U = u) \\ \text{s.t.} &\frac{(\text{eCPI}_{max}(A) - \text{bid}_a \cdot P(C_a = 1|U = u))}{\text{eCPI}_{max}(A)} \leq \tau \end{aligned} \tag{4}$$

Given Equation 4 we define *the controlled quality discounted eCPI*

$$\text{eCPI}_c(A) = \text{bid}_{a_c} \cdot P(C_{a_c} = 1|U = u)$$

Substituting regular quality discounted eCPI for controlled quality discounted eCPI and normalizing by the total revenue for eCPI_{max} in Equation 3 we obtain:

$$\frac{\sum_{j \in I_q \setminus I} \text{eCPI}_c(A_j)}{\sum_{i \in I_{(t,t_k)}} \text{eCPI}_{max}(A_i)} \geq \tau \tag{5}$$

By controlling the maximum loss in revenue on a per impression basis we limit the potential revenue loss to a fixed percentage. This allows publishers to predetermine the amount of risk they wish to take in exploring whether introducing quality ads in a market place will yield an increase in revenue over a standard serving model.

7.4 Impact of ad feedback filtering on revenue

As demonstrated in Figure 5, it may be costly to evaluate the revenue impact of showing users quality ads as it is unknown when (or if) the initial revenue investment (showing high quality but less monetizable ads) is offset by an increase in user engagement [15]. To investigate whether unbiased ad feedback estimates provide additional value over biased estimates we evaluate their effectiveness in reducing hide rate at different levels of short-term revenue investment. We again extract two weeks of ad interaction and revenue data from the logs of the Internet company. For each first slot auction we find the top 10 ads that participated based on $eCPI_{max}$.

We follow a simulation based approach based on [2] and simulate for each impression the ad that would have been shown using either a biased estimate (i.e., the full model from Section 6) or a bias corrected estimate (i.e., only the ad coefficients from Section 6), as well as the true hide rate for each auction using Equation 4.

Figure 6 shows the proportional revenue loss threshold, i.e., the percentage short term revenue investment on the x-axis (defined as τ in Section 7.3) and the reduction in ad feedback for each method at each threshold on the y-axis. For each threshold $\tau \in \{1, 5, 10\}$ percent short term revenue investment, we determine the ads shown in the simulated auctions for each method and find the reduction in hide rate compared to the original auctions.

We observe that at 1% short term revenue investment there is little difference between the biased and unbiased curves, implying that the initial ads that are filtered are found to be bad by all users. This is an important finding, as it shows that for the worst of the bad ads, ad feedback data is not biased and can be used as is to remove bad ads. Furthermore, the ads (and their hides) that are removed by the bias corrected estimates are received from all users, suggesting they are considered of low quality by the general population. While ads removed based on the original quality estimates may have been received from only a particular sub-population.

We observe the true benefit of unbiased ad feedback estimates at a 10% revenue investment as at that point a 20% reduction in hides can be achieved compared to about a 12% reduction using biased estimates. Moreover, at 10% the short term revenue investment is far lower than the 28% to 37% cost when using the traditional quality based ranking approach while achieving higher quality improvements, cf. Figure 5. Based on business criteria other trade-offs may be selected; however, our results suggest that there is value

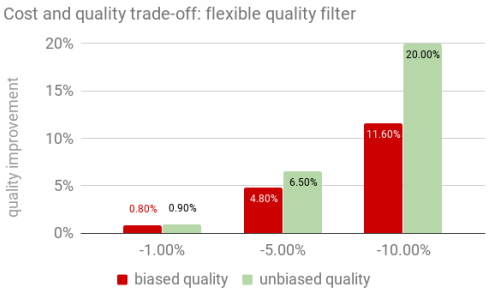


Figure 6: Revenue investment-hide reduction trade-off as rate of total when filtering ads using eqn. 4 based on biased (red), and unbiased (green) estimates of hide rate.

in using unbiased estimates in that they provide the best return on investment.

8 CONCLUSIONS AND FUTURE WORK

Publishers relying on advertising for revenue are aware that optimizing for CTR is not sufficient for long-term engagement. They must also consider the quality of the ads shown to their users. To this end, some provide mechanisms to gather explicit feedback about the ads. This work is the first providing insights into the sources of bias present in ad feedback data, i.e. hide rate, in estimating ad quality.

Using hide rate, we focused on understanding and modeling the characteristics of users who do not want to see ads. An initial analysis of the ad feedback data shows that some types of users may be exposed to lower quality ads than others and as such result in higher feedback rates. More precisely, we found evidence for the presence of two sources of selection bias, i.e., targeting and response bias. To correct for such bias we develop a so-called user selection bias model that allowed us to quantify the bias present in the feedback on ads, considering both users' characteristics (ad targeting criteria, such as age and interests) and their behaviour (engagement levels).

We then incorporate the user selection bias model into an ad quality model, which results in a bias corrected ad model that provides a true estimate of the inherent quality of the ads. The corrected estimate is the unbiased hide rate of an ad that controls for the type of users that ad has been shown. Further, we introduced a method to control revenue investment when using ad feedback estimates in practice. We found that when comparing the use of biased ad feedback estimates with bias corrected ones in a business scenario that unbiased estimate provide a benefit in terms of the quality-revenue trade-off. However, these benefits only materialize at certain levels of revenue investment and with a mechanism that is able to carefully control revenue investment. Our work allows Internet companies to experiment with ad quality at a fixed risk level, lowering the threshold to explore further quality signals as well as ranking techniques to improve users' experience with ads.

We acknowledge that our findings are based on a sample of data from one Internet company, Yahoo news app. However, the size of our sample and the use of features in our analyses are generally available in the advertisement industry, therefore providing ad networks and publishers with the insights necessary to mitigate user bias in ad feedback data.

REFERENCES

[1] M. Aharon, N. Aizenberg, E. Bortnikov, R. Lempel, R. Adadi, T. Benyamini, L. Levin, R. Roth, and O. Serfaty. Off-set: one-pass factorization of feature sets for online recommendation in persistent cold start settings. In *RECSYS'13*, pages 375–378. ACM, 2013.

[2] A. Ashkan and C. L. A. Clarke. Modeling browsing behavior for click analysis in sponsored search. In *CIKM*, 2012.

[3] S. Athey and D. Nekipelov. A structural model of sponsored search advertising auctions. In *Sixth ad auctions workshop*, volume 15, 2010.

[4] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern. Visual appearance of display ads and its effect on click through rate. In *CIKM*, 2012.

[5] N. Barbieri, F. Silvestri, and M. Lalmas. Improving post-click user engagement on native ads via survival analysis. In *WWW'16*, pages 761–770, 2016.

[6] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: Quantifying the impact of landing pages in web advertising. In *CIKM'09*, 2009.

[7] G. Brajnik and S. Gabrielli. A review of online advertising effects on the user experience. *IJHCI*, 26(10):971–997, 2010.

[8] A. Z. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: learning when (not) to advertise. In *CIKM'08*, 2008.

[9] J. M. Chambers and T. J. Hastie. *Statistical models in S.* CRC Press, Inc., 1991.

[10] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *SIGKDD'12*, pages 777–785. ACM, 2012.

[11] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang. Using landing pages for sponsored search ad selection. In *WWW'10*, 2010.

[12] H. Cramer. Effects of ad quality & content-relevance on perceived content quality. In *CHI '15*, pages 2231–2234, New York, NY, USA, 2015. ACM.

[13] D. G. Goldstein, R. P. McAfee, and S. Suri. The cost of annoying ads. In *WWW'13*, 2013.

[14] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, R. Baeza-Yates, A. Feng, E. Ordentlich, L. Yang, and G. Owens. Scalable semantic matching of queries to ads in sponsored search advertising. In *SIGIR'16*, pages 375–384, 2016.

[15] H. Hohnhold, D. O'Brien, and D. Tang. Focusing on the long-term: It's good for users and business. In *KDD'15*, pages 1849–1858, 2015.

[16] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression.* John Wiley & Sons, 2004.

[17] S. Ieong, M. Mahdian, and S. Vassilvitskii. Advertising in a stream. In *WWW'14*, 2014.

[18] A. Kae, K. Kan, V. K. Narayanan, and D. Yankov. Categorization of display ads using image and landing page features. In *LDMTA*, 2011.

[19] D. Kempe and B. Lucier. User satisfaction in competitive sponsored search. In *WWW'14*, 2014.

[20] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users.

[21] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users. In *SIGKDD'15*, pages 1929–1938, 2015.

[22] R. J. Oentaryo, E.-P. Lim, J.-W. Low, D. Lo, and M. Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *WSDM'14*, 2014.

[23] R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *WSDM'12*, 2012.

[24] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *KDD'09*, 2009.

[25] G. Shmueli. To explain or to predict? *Statistical science*, pages 289–310, 2010.

[26] E. Sodomka, S. Lahaie, and D. Hillard. A predictive model for advertiser value-per-click in sponsored search. In *WWW'13*, 2013.

[27] F. Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169, 1998.

[28] T. Wang, J. Bian, S. Liu, Y. Zhang, and T.-Y. Liu. Psychological advertising: exploring user psychology for click prediction in sponsored search. In *SIGKDD'13*, 2013.

[29] D. Yin, B. Cao, J.-T. Sun, and B. D. D. 0001. Estimating ad group performance in sponsored search. In *WSDM'14*, 2014.

[30] K. Zhou, M. Redi, A. Haines, and M. Lalmas. Predicting pre-click quality for native advertisements. In *WWW'16*, pages 299–310, 2016.

Implicit Bias in Crowdsourced Knowledge Graphs

Gianluca Demartini
University of Queensland, Australia
demartini@acm.org

ABSTRACT

Collaborative creation of knowledge is an approach which has been successfully demonstrated by crowdsourcing project like Wikipedia. Similar techniques have recently been adopted for the creation of collaboratively generated Knowledge Graphs like, for example, Wikidata. While such an approach enables the creation of high quality structured content, it also comes with the challenge of introducing contributors' implicit bias in the generated Knowledge Graph. In this paper, we investigate how paid crowdsourcing can be used to understand contributor bias for controversial facts to be included into collaborative Knowledge Graphs. We propose methods to trace the provenance of crowdsourced fact checking thus enabling bias transparency rather than aiming at eliminating bias from the Knowledge Graph.

ACM Reference Format:
Gianluca Demartini. 2019. Implicit Bias in Crowdsourced Knowledge Graphs. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317307

1 INTRODUCTION

General-purpose Knowledge Graphs incorporate factual statements that can be accessed by end-users and that can be integrated across different structured datasets [3]. Knowledge Graphs are typically constructed in a bottom-up fashion by making semantics emerge from data like, for example, textual document collections or Web pages [2, 12]. However, Knowledge Graphs can also be created by means of large-scale manual effort like for example the classic Cyc ontology project [13] that required 900 person-years to be created. More recently, the crowdsourcing project Wikidata [21] was kicked-off to generate a general purpose Knowledge Graph by collecting input from thousands of individual contributors who continuously add and edit factual statements in the Knowledge Graph.

The use of crowdsourcing in combination with machine-based algorithms has been often used for different Semantic Web problems [8, 14, 19]. This is typically done thanks to crowdsourcing platforms like Amazon MTurk [6] which enable programmatic access to a crowd of contributors by means of Human Intelligence Tasks (HITs) that are published by requesters and completed by workers in exchange for small monetary rewards. Common crowdsourcing tasks in the Semantic Web domain include schema mapping, entity extraction and linking, relation extraction, and ontology construction and verification.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317307

While leveraging crowdsourcing for Knowledge Graph construction, human contributors bring their own point of view on the factual statements they add to the Knowledge Graph. For controversial statements like, for example, Catalonia being part of Spain or being an independent country, contributors may bring their own implicit bias into the Knowledge Graph. There are different ways to deal with this issue. On the one hand, it is possible to pre-select crowd contributors and assign them editorial micro-tasks based on user profiles (e.g., including information on contributors' background) and to build a sample of contributors which is representative of the population, thus removing or diminishing any possible bias effect. On the other hand, we can maintain the idea of having an open call for contributions as in classic crowdsourcing projects, but then keep track of information about workers' background as well as about the information seeking process (i.e., which search engine has been used to retrieve the information, which search query, and which result has been used to identify the contributed information) used to reach a certain conclusion thus tracking potential bias in the answers.

In this paper we focus on the latter approach to deal with bias in crowdsourced Knowledge Graphs: We propose a model to measure and track the implicit bias of crowd contributors into the Knowledge Graph. We present the results of a study of crowdsourced controversial fact verification asking the crowd to provide evidence supporting their claims. Based on such supporting evidence we are able to identify the different points of view that exist in the crowd and keep track of bias in the Knowledge Graph together with the factual statements inserted into it.

The main contribution of this paper are the following:

- We perform a large-scale study involving 600 crowd workers to understand what aspects of worker demographics and of the knowledge creation process significantly affect the statements which are stored in Knowledge Graphs.
- We propose a model to keep track of bias information in crowdsourced Knowledge Graphs like Wikidata and demonstrate the benefits of surfacing bias information to end users of applications like, for example, Semantic Search.

The rest of this paper is structured as follows. In the next section we briefly summarize related work about bias in Web Search and on the use of crowdsourcing for Semantic Web data. Then, in Section 3 we present the results of three crowdsourcing studies aimed at understanding which are the dimensions that significantly affect the answer to controversial fact checking HITs to decide what bias information is most useful to keep track of in Knowledge Graphs. In Section 4 we explain how such bias information can be stored in Knowledge Graphs. In Section 5 we discuss the implication of such additional bias data collection for storage and retrieval in Knowledge Graphs like Wikidata. Finally, in Section 6 we draw the main conclusions of this work.

2 RELATED WORK

Bias is a well studied concept in Psychology research. A popular example of a large-scale Web study of bias is the Implicit Bias project run by Harvard [9] where reaction tests have been performed to measure implicit bias towards gender, skin color, and other dimensions.

In the area of crowdsourcing, bias has been studied less, but some previous work serves as the basis for this study. Some work has studied how crowdsourcing contributors are biased towards certain attitudes. For example, Otterbacher [16] looked at how certain images lead to diverse annotations in a game-with-a-purpose setting. Results show that images depicting women tend to be described using more subjective adjectives thus confirming the presence of gender-based stereotypes.

Eickhoff [7] looked at the effect of cognitive bias in crowdsourcing. He showed how crowd workers are affected by fellow workers' answer (Bandwagon effect) and by being presented with multiple options (Decoy effect). The existence of the Decoy effect proves that workers judgment is indeed affected by other tasks they have seen before approaching a specific HIT and, more generally, by the their cultural background defined as all the information they have been exposed to before reaching the specific crowdsourcing task. Hube et al. [10] measured the effect of implicit bias in paid crowdsourcing focusing on subjective data labelling tasks. Roitero et al. [17] used crowdsourcing for fact checking presenting an experimental comparison of different judgment scales. In our work we look at the effect of crowd workers' implicit bias on the verification of controversial factual knowledge to be included in a Knowledge Graph.

Crowdsourcing platforms like Amazon MTurk enable access to workers having a large diversity of cultural backgrounds [18]. The two countries where most crowd workers on these platform are from are United States and India [6]. In this paper we focus on these two different groups also because of the different cultural context they are embedded into and how it affects their answers on controversial fact checking HITs.

A common approach to deal with noise and low quality data in crowdsourcing is to aggregate answers collected for the same task from multiple workers by, for example, majority vote or by using more advanced aggregation methods [20]. An alternative approach to aggregating answers is to analyze information about disagreement among workers [1] or to compute confidence measures about the answer [11] by means of appropriate agreement measures [4].

Crowdsourcing has been extensively used for Semantic Web and Linked Data problems. Popular examples of human-in-the-loop solutions using paid micro-task crowdsourcing include the work on schema mapping by Sarasua et al. [19]. Another example demonstrates the use of games-with-a-purpose for entity extraction in text [8]. Mortensen et al. [14] show how crowdsourcing can be used to verify relationships among classes of domain-specific ontologies. They show that even in cases where domain expertise is required to complete the task accurately, crowd workers can perform as good as experts if provided with enough contextual information about the classes they are considering in the HIT. For a comprehensive overview of human-in-the-loop solutions for Web Science problems we refer to [5]. The question we aim to answer in this work is on the effect of crowd workers' implicit bias contributing data to Knowledge Graphs. A strong motivational example for this research question is Wikidata [21] which is a collaboratively created Knowledge Graph to which anyone on the Web can contribute new facts and modify existing ones.

Tracking provenance information in RDF is not a new problem but early solutions have shown to be inefficient by creating an overhead in terms of data to be stored and in indexing resourced required [15]. However, more recently, RDF stores that can efficiently index and query provenance metadata have been created [23]. More than better RDF stores, better ways to represent meta-information about RDF statements (which are more efficient than reification) are being used in practice in Knowledge Graphs like YAGO2 [15]. Such work enables the methods described in our paper by providing a scalable and reliable infrastructure for bias-tracking in Knowledge Graphs.

3 CROWDSOURCING CONTROVERSIAL FACT CHECKING

In this section we report the results of an experimental study we performed on top of the Amazon MTurk crowdsourcing platform about how crowd workers can contribute to the verification of controversial facts and on what factors influence their contributions.

3.1 Crowdsourcing Setting

We analyzed three different scenarios including two controversial fact verification tasks and one fake news verification task. For each use case, we recruited 200 crowd workers (100 from US and 100 from India) for a total of 600 workers involved in the study run in January 2018. We rewarded them $0.05 for controversial fact HITs and $0.10 for the fake news verification HIT. For each task we asked a binary question (yes/no or true/false) about a certain statement or news video. For each answer they provided, we also asked workers to justify it: We requested to provide information about the search engine used to find the answer, which search query they used, which URL contains information in support to their answer, and in which position the URL was ranked in the search engine result page (SERP). We additionally collect information about workers' age and gender.

3.2 Study I: Catalonia Independence

The first controversial fact HIT we run asked crowd workers to answer the question and provide evidence on whether Catalonia is an independent country or not. We hypothesize this being a controversial fact given the independence referendum held in October 2017 in Catalonia which has been deemed unlawful by Spanish courts. Out of 200 workers, 116 answered 'yes' and 84 answered 'no' to the question on Catalonia independence, showing how the question is indeed controversial and a clear truth is not obviously obtainable by aggregating such answers. This is consistent with claims from previous work [1] in support of not aggregating crowd answers but rather maintaining the diversity of opinions expressed by the crowd. When comparing the two crowd populations (US and India) we observe that in both cases the majority of respondents answered 'yes' but with a smaller majority in US (47 vs 46) as compared to

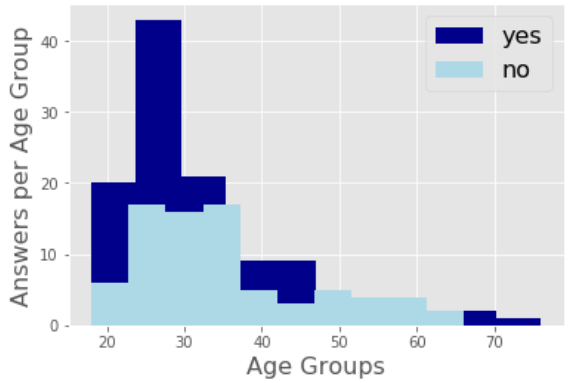


Figure 1: Age distribution of workers responding 'yes' and 'no' to the question on whether Catalonia is an independent country.

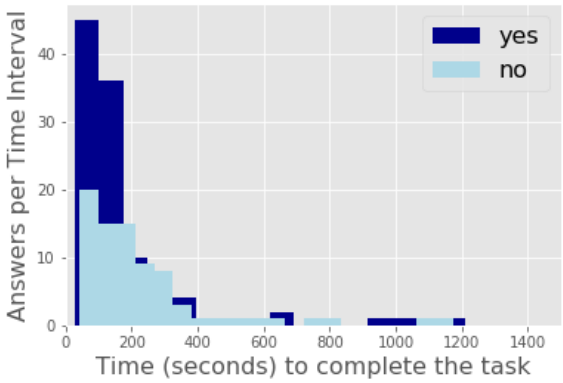


Figure 2: Task completion time distribution for workers responding 'yes' and 'no' to the question on whether Catalonia is an independent country.

India (62 vs 33). A Chi-squared test shows a not significant effect of the location of workers (p-value=0.058).

Age Bias. The age distribution of respondents is presented in Figure 1. We can see that a large group of workers who provided positive answers towards Catalonia independence tends to be younger than average. Indeed, the median age of workers who responded 'yes' is 28 as compared to the median age of those who responded 'no' which is 32.5. The effect of age on the answer for this HIT is significant (t-test $p < 0.05$) indicating that workers who answered 'yes' are significantly younger. This is another demonstration that based on the crowd population who provides factual statements to a Knowledge Graph, different information could be encoded in it.

Gender Bias. When looking at the distribution of answers over different gender groups we see that 35 male and 73 female workers answered 'yes' while 46 male and 33 female workers answered 'no'. We can observe a clear pattern in terms of gender bias among the respondents who provided their gender information: Female contributors achieved a strong majority for 'yes' while male contributors had a majority for 'no'. A Chi-square test shows a significant effect ($p < 0.01$) of gender on the answer.

Task Duration Bias. Figure 2 shows the distribution of the time taken to answer the questions in the task. In terms of time taken to complete the task, workers who responded 'yes' took a median time of 110 seconds to complete the task, while workers who answered 'no' took longer (median=180s). The difference in time is not significant showing that the time taken to complete the task had no impact on the actual answer.

Search Result Rank Bias. Figure 3 shows the ranking of the URL used to support the answer about Catalonia being an independent country as presented in the SERP used to identify the answer and relative supporting evidence. In terms of ranking, we can observe that the median ranking of URLs that support 'yes' answers is 1.62 which is significantly different (t-test $p < 0.01$) from the rank of URLs used as support by workers who answered 'no' (median=2.32).

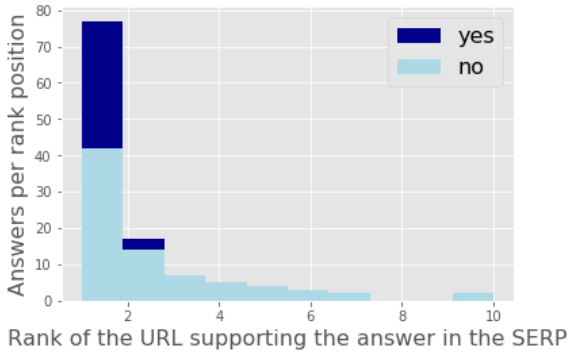


Figure 3: Search result ranking of the URL used as supporting evidence by workers responding 'yes' and 'no' to the question on whether Catalonia is an independent country.

This indicates that workers who go lower in the SERP to find the supporting evidence tend to answer 'no'.

Top-level Domain of Supporting Evidence. When looking at the most used domains for the supporting evidence URLs, we observe that the most popular sources of evidence are 'en.wikipedia.org ' (49 times), 'www.independent.co.uk' (43), 'www.google.com'[1] (20), 'www.telegraph.co.uk' (13), and 'www.bbc.com' (10). When differentiating the source of supporting evidence for 'yes' and 'no' answers, we obtain the results presented in Table 1. We can observe how the top domains for each answer are quite different showing a prevalence of 'www.independent.co.uk' for workers who answered 'yes'.

3.3 Study II: Israel Capital

A second experiment we run on Amazon MTurk asked 200 workers (100 from US and 100 from India) which is the capital city of Israel

[1]The URLs having as domain www.google.com are SERPs indicating how workers picked their answers based on the snippets of information provided in the SERP without clicking further away to any retrieved result.

Table 1: Top-3 domains of URLs supporting 'yes' and 'no' answers to whether Catalonia is an independent country.

Domain supporting Yes	Count	Domain supporting No	Count
www.independent.co.uk	35	en.wikipedia.org	26
en.wikipedia.org	23	www.bbc.com	10
www.google.com	18	www.telegraph.co.uk	9

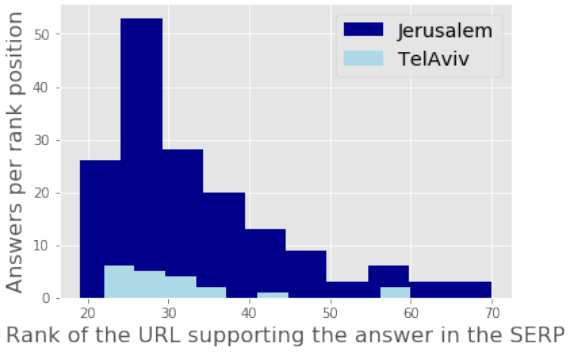


Figure 4: Age distribution of workers responding to the question on which is the capital of Israel.

giving the options 'Jerusalem', 'Tel Aviv', and 'Other' as possible answers. We hypothesize this being a controversial fact given the decision to move the US embassy from Tel Aviv to Jerusalem and the following discussion that took place at the United Nations. Out of 200 workers, 164 answered 'Jerusalem', 20 answered 'Tel Aviv', and 10 answered 'Other' to the question, showing how the question is less controversial than the previous one about Catalonia independence. Even in this case we can perform an analysis of the supporting evidence provided by the crowd workers to understand where the minority point of view comes from and aim at not discarding it by simply aggregating the data and taking the most popular answer.

When comparing workers from India and US we did not observe a significant difference (chi-square $p > 0.05$) in their answers with a strong majority for 'Jerusalem' as the answer in both populations.

Age Bias. The age distribution of respondents is presented in Figure 4. We can see that the distribution of the age from the two groups is similar with median values of 30 for the group answering 'Jerusalem' and 28.5 for the group answering 'Tel Aviv'. No statistically significant effect of age has been observed.

Gender Bias. Figure 5 shows the gender bias in the answers. Again, we can see similar distributions across the two groups and no significant effect of gender on the answer has been observed.

Task Duration Bias. Figure 6 shows the distribution of the time taken to answer the questions in the task. In terms of time taken to complete the task, we can observe a similar distribution for both answers (the median time of 'Jerusalem' respondent was 140.5s; the median time of 'Tel Aviv' respondent was 143.5s). Indeed, a statistical test shows no significant differences in the time taken to answer.

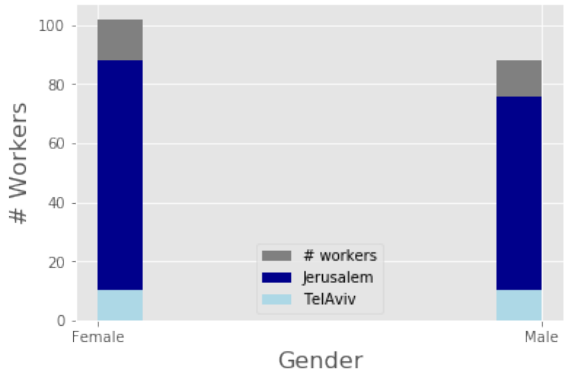


Figure 5: Gender distribution of workers responding to the question on which is the capital of Israel.

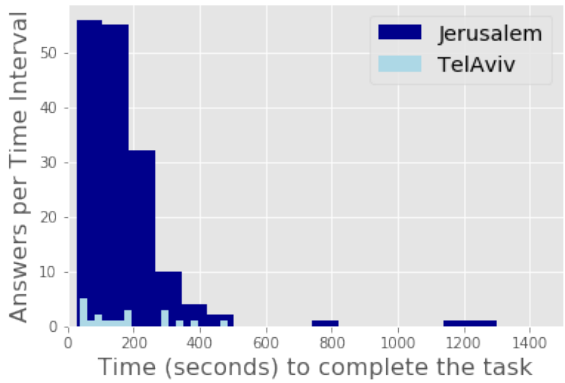


Figure 6: Task completion time for workers responding to which is the capital of Israel.

Search Result Rank Bias. Figure 7 shows the ranking of the URL used to support the answer about the capital of Israel as presented in the SERP used to identify the answer and the relative supporting evidence. In terms of ranking, we can observe very similar patterns with the median rank of supporting URLs at 2.3 for 'Jerusalem' and at 2.35 for 'Tel Aviv' with no statistically significant difference.

3.4 Study III: Pope News

Finally, we crowdsourced a task where we asked crowd workers whether a video[2] showing the Pope performing a trick was real or fake, again, also asking to provide supporting evidence for their claim. In this case the fact was not controversial but rather meant to check the veracity of the news which was not real but made up for a TV show.

Out of the 200 crowd workers involved in the study, 107 said that the video news story is real while 93 said that it is fake. In this case, despite the objectivity of the fact being false, we observed high disagreement levels.

[2] http://www.youtube.com/embed/MYjeWGH8Z1c

When comparing the two crowd population (US and India) we observe that the majority of Indian workers believed the news being real (66 vs 34) while the majority of workers from US (59 vs 41) believed the news was fake. The fake video was indeed created for a US-based TV show justifying the higher awareness of US workers. A Chi-squared test shows a significant difference among the answering pattern of the Indian and US-based worker populations ($p < 0.01$).

Age Bias. Figure 8a shows the distribution of workers' age based on their answer to the fake news identification task. We can observe that workers who believe the news being authentic are younger (median=28) as compared to those believing the news being fake (median=32). A t-test shows a statistically significant effect of workers' age on believing or not to the news ($p < 0.01$) demonstrating how younger contributors are more prone to being mislead by this fake news.

Task Duration Bias. Figure 8b shows the distribution of the time taken to complete the HIT. While the median time to answer the task was higher (267.5s) for contributors who stated that news is not real as compared to contributors who said the news is real (187.5s) a t-test shows no significant differences on the time taken to answer based on the answer.

Search Result Rank Bias. Figure 9 shows the position of the supporting evidence identified by workers in the SERP for the fake news identification task. The median rank of the supporting evidence for respondents who believe the news is real was 2 while for respondent who believe the news is fake was 2.3. While workers who did go deeper in the ranking managed to identify the non-truthfulness of the news, no statistically significant difference has been observed based on the rank of the evidence.

3.5 Result Discussion

From the results presented above we understand that while certain facts may be more controversial than others (with the 'Catalonia independence' task showing more controversy than the 'Israel capital'

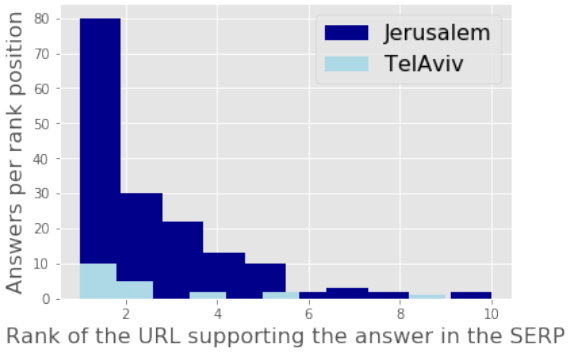


Figure 7: Search result ranking of the URL used as supporting evidence by workers responding to the question on which is the capital of Israel.

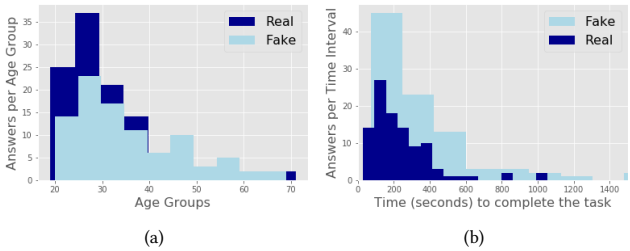


Figure 8: (a) Age of crowd workers responding to the task of fake news identification. (b) Time taken to answer to the fake news task by crowd workers.

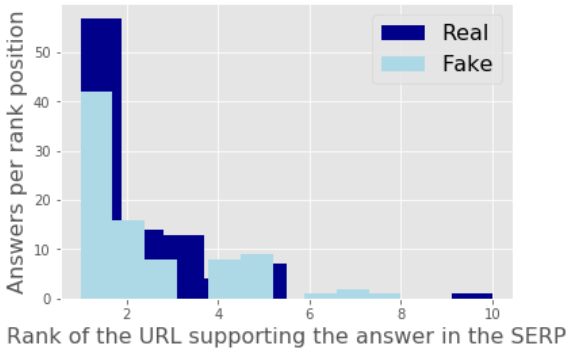


Figure 9: Search result ranking of the URL used as supporting evidence by workers responding to the fake news identification task.

task), we have observed the importance of keeping track of information provenance (both contributors' background and the source of evidence) when creating a crowdsourced Knowledge Graph.

A recent attempt to keep track of provenance in Knowledge Graphs is the concept of *reference* in Wikidata. For example, the statement 'Tom Cruise' - 'date of birth' -'3 July 1962' has 6 references among which one is from English Wikipedia[3]. While this is useful information to support facts stored in Knowledge Graphs, it lacks meta-information about how this was identified as a source. In the experiments we presented above, we have shown the effect of how different contributors may end up providing different statements by using different support evidences as reference for the statement. We claim it is then important to incorporate provenance metadata about contributors' implicit bias as well as the statement source of evidence.

4 TRACKING BIAS INFORMATION IN KNOWLEDGE GRAPHS

In this section we present a model to track bias information for controversial statements in Knowledge Graphs by adding annotations about information provenance and about how supporting evidence has been identified by contributors. Specifically, we propose to give

[3] https://www.wikidata.org/wiki/Q37079

Figure 10: Knowledge Graph entity with controversial facts.

the option for each fact stored in a crowdsourced Knowledge Graph like Wikidata to keep information about the Search Engine used to retrieve the reference supporting the fact, the search query used to retrieve the reference, and the position in the ranked list of results in the SERP in which the reference was found.

Such additional metadata can be stored in Knowledge Graphs using existing techniques like, for example, *reification* in RDF [15]. This allows us to generate statements about statements thus making it possible to provide references for statements (e.g., like in Wikidata) and to annotate subgraphs with metadata about them (e.g., supporting URLs). While reification has been shown to generate a large volume of data if used extensively, alternative approaches to deal with the overhead introduced by reification also exist. For example, in [22] authors assign temporal validity information to facts by means of custom properties to create additional statements about statements:

- #3: DavidBeckham playsFor RealMadrid
- #3 since 2003
- #3 until 2007

In a similar way we propose to trck provenance and information about implicit bias which can be efficiently stored and retrieved from custom knowledge bases like, for example, TripleProv [23].

In the case of the Catalonia use case, we would store the following statements:

- #1: Catalonia instanceOf 'autonomous community of Spain'
- #1 supportedBy 42%
- #1 mainSourceOfEvidence en.wikipedia.org
- #1 avgRankOfEvidence 2.32
- #1 medianAgeOfContributors 32.5
- #2: Catalonia instanceOf 'sovereign state'
- #2 since 2017
- #2 supportedBy 58%
- #2 mainSourceOfEvidence independent.co.uk
- #2 avgRankOfEvidence 1.62
- #2 medianAgeOfContributors 28

5 DISCUSSION

Modern data models for Knowledge Graphs like the one used by Wikidata already allow for the use of qualifiers (e.g., start and end

Figure 11: Entity card in SERP proposing to the end user alternative statements from the background Knowledge Graph with bias information.

date of the validity of a statement like 'President of the United States') and references for statements[4]. The model we propose builds on top of existing data models and existing RDF stores and allows to embed in the Knowledge Graph information about the bias certain statements carry with them from the data creators. Such provenance information can then allow data consumers to be aware of possible existing bias in the data and to deal with it as they deem appropriate (e.g., selecting the majority answer or selecting the answer preferred by the demographics group they belong to).

As the information stored in Knowledge Graphs can be highly influenced by the population of contributors that contributed it, such information would be biased and follow up usage of the Knowledge Graphs (e.g., empowering Semantic Search systems) could lead to biased results returned to end users. Thus, rather than avoiding to store biased information in the Knowledge Graph, we believe it is a better option to keep track of bias (see Figure Figure 10 for an example) in order to better inform end users of potential bias and envision such information to be surfaced to them as, for example, shown in Figure 11 for entity cards in SERPs.

6 CONCLUSIONS

In this paper we discussed the need to track bias information into Knowledge Graphs. We presented the results of three crowdsourcing experiments involving 600 crowd workers that show how factual statements stored into Knowledge Graphs can be controversial. We have motivated the need for keeping track of provenance metadata for such controversial statements and shown how different populations of contributors can significantly influence the facts stored in a Knowledge Graphs.

[4] https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer

Future directions for this line of research include the evaluation with end users of the benefits of surfacing bias and provenance information in applications like Semantic Search. More than this, it is important to design automatic methods for the selection of the subset of statements in a crowdsourced Knowledge Graph that would benefit most from bias metadata being preserved due, for example, to their controversial dimension.

Acknowledgments. This project is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 732328.

REFERENCES

[1] Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine 36(1), 15–24 (2015)
[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The semantic web pp. 722–735 (2007)
[3] Bizer, C.: The emerging web of linked data. IEEE intelligent systems 24(5) (2009)
[4] Checco, A., Roitero, A., Maddalena, E., Mizzaro, S., Demartini, G.: Let's agree to disagree: Fixing agreement measures for crowdsourcing. In: 5th AAAI Conference on Human Computation and Crowdsourcing (HCOMP). pp. 11–20 (2017)
[5] Demartini, G., Difallah, D.E., Gadiraju, U., Catasta, M.: An introduction to hybrid human-machine information systems. Foundations and Trends in Web Science 7(1), 1–87 (2017), https://doi.org/10.1561/1800000025
[6] Difallah, D.E., Catasta, M., Demartini, G., Ipeirotis, P.G., Cudré-Mauroux, P.: The dynamics of micro-task crowdsourcing: The case of amazon mturk. In: 24th International Conference on World Wide Web. pp. 238–247. WWW '15 (2015)
[7] Eickhoff, C.: Cognitive biases in crowdsourcing. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 162–170. ACM (2018)
[8] Feyisetan, O., Luczak-Rösch, M., Simperl, E., Tinati, R., Shadbolt, N.: Towards hybrid NER: A study of content and crowdsourcing-related performance factors. In: 12th European Semantic Web Conference, ESWC. pp. 525–540 (2015)
[9] Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology 74(6), 1464 (1998)
[10] Hube, C., Fetahu, B., Gadiraju, U.: Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Proceedings of the 37th Annual ACM Conference on Human Factors in Computing Systems. CHI '19 (2019)
[11] Maddalena, E., Roitero, K., Demartini, G., Mizzaro, S.: Considering assessor agreement in ir evaluation. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. pp. 75–82. ICTIR, ACM (2017)
[12] Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent systems 16(2), 72–79 (2001)
[13] Matuszek, C., Cabral, J., Witbrock, M.J., DeOliveira, J.: An introduction to the syntax and content of cyc. In: AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering. pp. 44–49 (2006)
[14] Mortensen, J.M., Musen, M.A., Noy, N.F.: Crowdsourcing the verification of relationships in biomedical ontologies. In: AMIA Annual Symposium Proceedings. vol. 2013, p. 1020. American Medical Informatics Association (2013)
[15] Nguyen, V., Bodenreider, O., Sheth, A.: Don't like rdf reification?: Making statements about statements using singleton property. In: 23rd International Conference on World Wide Web. pp. 759–770. WWW '14 (2014)
[16] Otterbacher, J.: Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1955–1964. CHI '15, ACM (2015)
[17] Roitero, K., Demartini, G., Mizzaro, S., Spina, D.: How many truth levels? six? one hundred? even more? validating truthfulness of statements via crowdsourcing. In: Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media (2018)
[18] Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: Shifting demographics in mechanical turk. In: CHI '10 Extended Abstracts on Human Factors in Computing Systems. pp. 2863–2872 (2010)
[19] Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. In: International Semantic Web Conference. pp. 525–541 (2012)
[20] Venanzi, M., Guiver, J., Kazai, G., Kohli, P., Shokouhi, M.: Community-based bayesian aggregation models for crowdsourcing. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 155–164. WWW '14, ACM (2014)
[21] Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
[22] Wang, Y., Zhu, M., Qu, L., Spaniol, M., Weikum, G.: Timely yago: Harvesting, querying, and visualizing temporal knowledge from wikipedia. In: 13th International Conference on Extending Database Technology. pp. 697–700. EDBT (2010)
[23] Wylot, M., Cudre-Mauroux, P., Groth, P.: Tripleprov: Efficient processing of lineage queries in a native rdf store. In: Proceedings of the 23rd international conference on World wide web. pp. 455–466. ACM (2014)

I'm Lonely. Who Should I Talk to?

Lauren Fratamico
Massachusetts Institute of Technology
fratamico@mit.edu

Deb Roy
Massachusetts Institute of Technology
dkroy@media.mit.edu

ABSTRACT

Loneliness is becoming a global epidemic. As many as 33% of Americans report being chronically lonely, with similar percentages being reported in countries around the world. Additionally, this is a percentage that has risen in recent years. Many are turning to online forums as a way to connect with others about their feelings of loneliness and to begin to reduce these feelings. However, oftentimes, posts go unresponded to and online conversations do not take place, perhaps because those conversing did not find a connection with each other, potentially leaving the poster feeling even more lonely. This paper introduces a human-in-the-loop approach so that computers can mediate interactions online about loneliness and facilitate more intimate interactions. We also discuss ways to mitigate the bias when creating this system. The artificial intelligence in this approach takes into account the homophilous characteristics of the conversations that are taking place online by examining the homophily of the participants. Initial findings related to the correlation between homophily and successful conversations about loneliness on reddit are presented, and we lay the groundwork for being able to facilitate finding optimal conversation partners for those that are feeling lonely by proposing a human-in-the-loop approach.

CCS CONCEPTS

• **Information systems → World Wide Web**; • **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

loneliness, depression, human-in-the-loop, homophily

ACM Reference Format:
Lauren Fratamico and Deb Roy. 2019. I'm Lonely. Who Should I Talk to?. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3317309

1 INTRODUCTION

Loneliness is a crippling epidemic around the world. Globally, as much as 40% of people are estimated to experience loneliness at some point in their lives, and this is a percentage that has doubled over the past 50 years. It can impact all ages, ranging from small children to the elderly. Due to isolation, many people go days without human contact, with an estimated 25% of adults over 75 going a month without seeing another person. Loneliness does not just manifest mentally, but can cause physical damage as well. These physical side-effects have been estimated to cost the US an additional 7 billion dollars in health care costs per year[1]. Other nations are beginning to see that this is a major public health crisis, so much so that the UK recently hired a minister of loneliness.

Face-to-face interactions are ideal for combating feelings of isolation, but this is not always possible due to shyness or medically-necessitated bed rest. As a result, many individuals turn to the internet to connect with others. One such example is Reddit, which is full of vibrant, supportive communities where people can converse, offering advice and connecting with people they have never met before in person. Research has shown that posting in forms like Reddit about depression can actually improve your mood over time, as indicated by both the language of a user's post becoming more positive [17] and the lexical diversity and readability improving over time [18].

On Reddit, people posts about all kinds of topics, as can be seen in Figure 1. People like to share stories, ask for advice, share positive updates, and tragically, there's even a few suicidal posts. Fortunately, many of these posts are not going into a vacuum, but are instead being commented on by others in the community, often resulting in vibrant back-and-forth conversation.

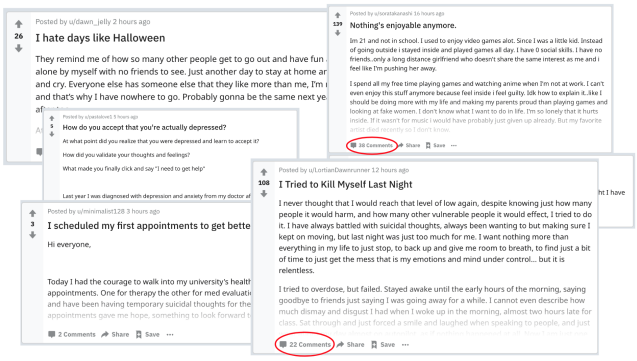


Figure 1: Example posts from Reddit on the topic of loneliness or depression. Post range from shared stories, advice seeking, positive updates, and suicidal posts.

In general, there are two types of comments to these posts, ones that result in conversation (where two or more people are engaging back and forth), and ones that do not (and instead get no interaction as if the commenter is posting into a void). An example of each of these types can be seen in Figure 2. As face-to-face interactions are the best for combating feelings of isolation, some form of interaction (even if it is just an interaction on an online forum) is better than no interaction.

[1] https://www.thecostofloneliness.org/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317309

In this paper, we propose a human-in-the-loop approach whose AI takes into account the characteristics of the individuals who are conversing. We use homophily to compute the similarity of the individuals conversing. Homophily is the idea that people who are similar prefer each other, which has been shown to be true in many network analyses [8, 14]. Humans are needed in this process to label which characteristics of similarity were the ones that drove the connection in the conversation. We aim to develop a system that finds the best person for a given individual to connect with (either online or through a helpline) so they can have the most useful conversation.

In this paper, we first discuss related work on homophily, then introduce our methods to compute homophily across those participating in online conversations about loneliness, then present initial results showing the promise of this approach, and then discuss how this would be contained within a human-in-the-loop approach to help people find the best conversation participant to maximize their chance of a useful conversation.

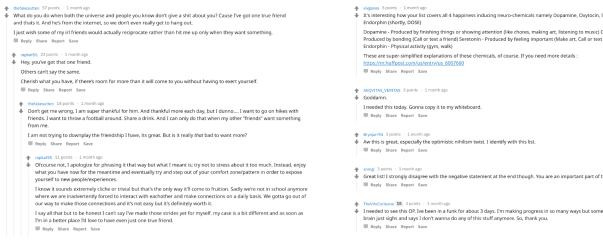


Figure 2: Two types of post interactions on Reddit. The left shows a positive interaction where two people are replying back and forth, and the right shows no interaction where many people replied to the post, but the original poster did not comment on any of them.

2 RELATED WORK

Homophily is the tendency of individuals to associate and bond with those that are like themselves, or, proverbially, that "birds of a feather flock together". Network analysis has consistently shown that this is the case, whether people are choosing adolescent friendships [8], romantic partners [9], or even doctors [7]. This is not just the case for in-person interactions, but in the online space as well, with researchers looking at friendships in Myspace [22] and on an online messaging platform [14].

Typical definitions of homophily use demographics to define the similarity between people, for example using race, sex, gender, or common language [2, 5, 11, 21]. While these are prominent markers of similarity between two people, they are harder to mine from a website like Reddit where people do not have profiles created for themselves that display this information as is the case on a website like Facebook. What can be mined from a site like Reddit is interests of a user, based on what subreddits they're posting on or what they have talked about in the past, and homophily can also be defined based on shared interests (as opposed to shared demographic qualities). It is evident that in-person friend groups tend to form based on these interest-based homophilous characteristics in addition to demographics-based ones [4], but on a forum like Reddit, people are interacting that do not know each other in person. In some cases, communities tend to form around common interests online, even if the people do not know each other. Chang et al. [3] studied activity on Pinterest and found that repinning (resharing) happened more amongst those that were interested in the same topic than those that were previously friends and following each other on Pinterest, showing that shared interests may be a stronger driver of activity than social connections. This indicates that communities do form online among those that have not previously interacted in real-life and that these communities are interest-driven. Communities do form on Reddit based on common interest, as is emphasized by the presence of subreddits (a subreddit being a community by definition). However, it is unclear if this is the case for Reddit conversations and leaves the question unanswered of if people communicate better on Reddit who don't know each other, but do have a number of interests in common. Ren et al. [19] tested theories about community attachment by forming groups on the MovieLens film recommendation site based on similarity of movie tastes. They found that people felt more attached to those that had similar tastes in movies as they did, and closer in general to the groups where people were grouped based on similarity. This is perhaps more evidence that those on Reddit that have stronger homophilous bonds would converse more. On the contrary, Bisgin et al. [1] found that interest-based homophily was not enough to construct new friendships on platforms like BlogCatalog, Last.fm, and LiveJournal. However, the Reddit users that we will be examining do not need to form friendships with each other, but instead just have healthy one-time conversations. This is a much lower bar than forming a friendship.

3 APPROACH

3.1 Data

Reddit[2] is a social news aggregation website where users share posts on a variety of topics, comment on these posts, and up/down vote all submitted content [12, 16]. Posts are self-categorized by their poster into one of thousands of 'subreddits'. Each subreddit is like a community, and they vary widely in topics, ranging from gaming to fitness to food. Each subreddit is run like a community with moderators to help those in the community adhere to the community's rules (eg, no vulgarity) [15]. While the site also features chatrooms where people can have realtime discussions with each other, the bulk of the activity is on the posts made within the subreddits [13].

In July 2015, Jason Baumgartner released a complete copy of Reddit available for public download. This contains over 1.7 billion Reddit posts and their comments, along with all available metadata (author, subreddit, position in comment tree, and other fields that are available through Reddit's API)[3]. Since then, the entire dataset has been uploaded to Google's BigQuery[4] and updated so that it

[2] reddit.com
[3] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/
[4] https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts and https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

Table 1: Total posts and comments on the top 20 represented subreddits.

Subreddit	Total Posts	Total Comments
depression	616,909	2,477,838
relationships	83,670	232,724
offmychest	67,881	232,724
AskReddit	64,029	1,444,043
SuicideWatch	55,185	31,2717
NoFap	48,403	333,965
relationship_advice	34,241	302,547
Advice	27,719	127,512
r4r	26,565	58,211
Anxiety	24,165	112,789
raisedbynarcissists	21,672	205,413
lonely	21,486	108,508
bipolar	20,844	151,664
dirtypenpals	19,346	7,242
stopdrinking	16,128	187,388
Drugs	14,405	242,143
mentalhealth	14,391	52,300
trees	14,132	148,368
Needafriend	13,278	45,637
teenagers	13,171	160,966

now contains a more complete version of Reddit posts [6] and all posts from Reddit's creation (in June 2005) through October 2018. Many researchers have begun to use this Reddit dataset to investigate a wide variety of questions, including examining online hate speech [20] and detecting sarcasm [10], but so far no one has used Reddit to investigate the health of discussions about loneliness.

To compose the dataset for my investigation, we pulled all posts that contained one of the following 5 words: lonely, loneliness, alone, depressed, and depression. This totals 2.4 million posts between June 2005 and October 2018. Additionally, we pulled all comments on those posts (over 24.7 million comments) and all previous posts of authors of those posts and comments. Posts come from 48,602 different subreddits. Postprocessing was necessary to ensure that this dataset does not include too many posts that do not address topics of loneliness and depression, eg, those that talk about going somewhere 'alone', but not necessarily *feeling* 'alone'. Table 1 shows the top 20 represented subreddits (by number of posts). Many of the posts come from mental health or relationship subreddits. It's also interesting to note the relative number of comments per post. r/AskReddit is a subreddit with a much larger following than the others, and, as a result, has a higher average number of comments per post.

3.2 Homophily

As a motivating example for why mining shared homophily characteristics could contribute positively to promoting healthy conversation between individuals, on one post where someone had shared coping strategies for dealing with loneliness, a conversation had started to emerge between two people. Overall, this conversation went well, with the two trading coping strategies back and forth. However, one individual mentioned that they like to smoke weed to cope. In response, the other participant became judgemental, scolding them for coping in this manner, and the conversation halted. In this case, the conversation would likely have been more healthy if both individuals shared the same opinions on marijuana.

In this work, homophily is computed as the number of overlapping subreddits that those participating in the conversation have participated in in the past. For example, if there are two people in the conversation, and participant 1 has previously posted in 5 subreddits, and participant 2 has previously posted in 10 subreddits, the homophily will be calculated as the number of subreddits that overlap. In this example, this would be a number between 1 and 5. 1 is a minimum because they have both posted in the subreddit that they are currently conversing on. We chose this relatively simple calculation of homophily to illustrate the power of this approach to begin to understand the relationship between interests of the conversation participants and success of the conversation.

3.3 Dyads

We aimed to understand how two people in a conversation were interacting, as this would mimic a typical therapy session or a friend reaching out to another. To do this, we took a subset of the data in which only two people were conversing back and forth. Additionally, we used only the conversation chains that were length three or longer (ie, participant 1 comments, participant 2 replies, then participant 1 replies back).

4 RESULTS

We present the initial results for our top represented subreddit, r/depression, to highlight the promise of this approach. We aimed to understand if length of the conversation chain was correlated with the homophily of the participants conversing. There was a correlation between these two variables [r = 0.0116, n =187,842, p = 1.38e-06]. This suggests that people who engaged in longer conversations were more likely to have a higher number of overlapping areas of interest, as calculated by the number of subreddits that they had both previously posted in. Similar positive correlations were found for other subreddits, indicating that homophily of interests may be a good indicator for predicting successful conversation. This would be the first step in out proposed AI solution to identifying optimal conversation partners.

5 PROPOSED HUMAN-IN-THE-LOOP INTERFACE

Because homophily and conversation length are correlated, this prompts our desire to create a system in which a user would be presented with the optimal conversation participant as they are using Reddit. For example, when someone comes on reddit and wants to reply to some subset of comments, they would be presented with an interface where the comments are presented in the sorted order of likelihood of successful interaction with the commenter. The one at the top would be the person that the AI has predicted they would have the most successful interaction with (based on homophily of interests), and the one at the bottom would be the one they are most predicted to have no interaction with. The system could eventually go even further to predict how positive the interaction would be (again based on homophily), and hide/show comments accordingly.

The artificial intelligence, as is, is able to determine the overlap in homophily and even, in the near future, predict the likelihood of interaction, but where a human would be useful in the loop is to help the system identify the most salient homophily characteristics that personally contribute to a more successful conversation or the lacking homophilous characteristics that may also indicate the outcome of the conversation. For example, in the story given above about the conversation going well until marijuana was brought up, had one participant known that the other was an active poster on r/trees (the marijuana subreddit), they may not have engaged in the conversation or have known that this is a topic they should have avoided if they had strong opinions against marijuana.

It is essential to have a human in the loop in this case to guide the AI towards the most salient homophilous characteristics that actually contribute to a successful conversation. If this human-in-the-loop system is presented to all people on reddit who want to comment on a given topic (but have a selection of people they could reply to), we could crowdsource the labeling of which homophilous characteristics are most important. Inherently, this also helps to reduce the bias of the crowd sourced approach because those that are actually conversing would be the ones selecting the homophilous characteristics.

Ideally, the system could also guide users towards items they should bring up in a conversation so that they can bond more throughout the conversation and have a more successful conversation. Here again, the human-in-the-loop system could present some homophilous characteristics of the conversation participants as they are conversing to help the human bring up even better topics while conversing. For example, if it was discovered that both participants post in the r/Berkeley subreddit, even though neither lives there, it could be inferred that at some point they both have (or have spent time there), and topics related to Berkeley could be brought up which facilitates the bonding process while conversing.

More broadly, an artificially intelligent solution that examines and uses the homophilous characteristics of all participants could be put in place when people are interacting over more than just the internet. This system could match and prompt conversation for those that are calling in to a suicide helpline or to people wanting to be seen by a therapist.

6 CONCLUSION

When someone is reaching out with intimate feelings, such as those that people are sharing on reddit forums about loneliness, and even more critically, the suicidal posts on these forums, it is essential for them to be able to connect to the person that is most able to help them. Oftentimes, we are solely connected to people based on availability. But with the power of data and information mining, it is possible to learn qualities of individuals and their preferences. Matching on any of these properties is better than random matching and would facilitate easier connections between people. Additionally, for each participant to have knowledge of these similarities, they could more quickly find common interests to converse about or allude to while replying to each other.

We have proposed a human-in-the-loop approach both towards matching those in lonely and depressed conditions with those that would most likely be able to engage in a meaningful conversation with them and towards facilitating better conversation through highlighting shared interests.

REFERENCES

[1] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. 2012. A study of homophily on social media. *World Wide Web* 15, 2 (2012), 213–232.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G Terveen. 2014. Specialization, homophily, and gender in a social curation site: findings from pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 674–686.

[4] Robert Crosnoe. 2000. Friendships in childhood and adolescence: The life course and new directions. *Social psychology quarterly* (2000), 377–391.

[5] Sergio Currarini, Matthew O Jackson, and Paolo Pin. 2009. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* 77, 4 (2009), 1003–1045.

[6] Devin Gaffney and J Nathan Matias. 2018. Caveat Emptor, Computational Social Science: Large-Scale Missing Data in a Widely-Published Reddit Corpus. *arXiv preprint arXiv:1803.05046* (2018).

[7] Geir Godager. 2012. Birds of a feather flock together: A study of doctor–patient matching. *Journal of health economics* 31, 1 (2012), 296–305.

[8] Jill V Hamm. 2000. Do birds of a feather flock together? The variable bases for African American, Asian American, and European American adolescents' selection of similar friends. *Developmental psychology* 36, 2 (2000), 209.

[9] Steven J Heine, Julie-Ann B Foster, and Roy Spina. 2009. Do birds of a feather universally flock together? Cultural variation in the similarity-attraction effect. *Asian Journal of Social Psychology* 12, 4 (2009), 247–258.

[10] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579* (2017).

[11] Lauri Kovanen, Kimmo Kaski, János Kertész, and Jari Saramäki. 2013. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences* (2013), 201307941.

[12] Alex Leavitt and Joshua A Clark. 2014. Upvoting hurricane Sandy: event-based news production processes on a social news site. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1495–1504.

[13] Alex Leavitt and John J Robinson. 2017. The Role of Information Visibility in Network Gatekeeping: Information Aggregation on Reddit during Crisis Events.. In *CSCW*. 1246–1261.

[14] Jure Leskovec and Eric Horvitz. 2008. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 915–924.

[15] Adrienne Massanari. [n. d.]. *Participatory Culture, Community, and Play. Chapter 2. Defining reddit.* Peter Lang, Bern, Switzerland. https://www.peterlang.com/view/9781454194866/xhtml/chapter002.xhtml

[16] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346.

[17] Albert Park and Mike Conway. 2017. Longitudinal changes in psychological states in online health community members: understanding the long-term effects of participating in an online depression community. *Journal of medical Internet research* 19, 3 (2017).

[18] Albert Park and Mike Conway. 2018. Harnessing Reddit to Understand the Written-Communication Challenges Experienced by Individuals With Mental Health Disorders: Analysis of Texts From Mental Health Communities. *Journal of medical Internet research* 20, 4 (2018).

[19] Yuqing Ren, F Maxwell Harper, Sara Drenner, Loren Terveen, Sara Kiesler, John Riedl, and Robert E Kraut. 2012. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Mis Quarterly* (2012), 841–864.

[20] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159* (2017).

[21] Wesley Shrum, Neil H Cheek Jr, and Saundra MacD. 1988. Friendship in school: Gender and racial homophily. *Sociology of Education* (1988), 227–239.

[22] Mike Thelwall. 2009. Homophily in myspace. *Journal of the American Society for Information Science and Technology* 60, 2 (2009), 219–231.

Putting the Human in the Time Series Analytics Loop

Shima Imani
University of California, Riverside
siman003@ucr.edu

Sara Alaee
University of California, Riverside
salae001@ucr.edu

Eamonn Keogh
University of California, Riverside
eamonn@cs.ucr.edu

ABSTRACT

Time series are one of the most common data types in nature. Given this fact, there are dozens of query-by-sketching/ query-by-example/ query-algebra systems proposed to allow users to search large time series collections. However, none of these systems have seen widespread adoption. We argue that there are two reasons why this is so. The first reason is that these systems are often complex and unintuitive, requiring the user to understand complex syntax/interfaces to construct high-quality queries. The second reason is less well appreciated. The expressiveness of most query-by-content systems is surprisingly limited. There are well defined, simple queries that cannot be answered by any current query-by-content system, even if it uses a state-of-the-art distance measure such as Dynamic Time Warping. In this work, we propose a natural language search mechanism for searching time series. We show that our system is expressive, intuitive, and requires little space and time overhead. Because our system is text-based, it can leverage decades of research text retrieval, including ideas such as relevance feedback. Moreover, we show that our system subsumes both motif/discord discovery and most existing query-by-content systems in the literature. We demonstrate the utility of our system with case studies in domains as diverse as animal motion studies, medicine and industry.

KEYWORDS

Time Series; Similarity Search; NLP

1 INTRODUCTION

One of the fundamental questions a data scientist can ask is, "*Are there examples of this pattern in my dataset?*"[35]. In some cases, the analyst may have a query pattern (query-by-example [9]) taken from a different dataset. In other cases, the analyst may be able to *draw* a pattern of interest (query-by-sketching [39]) or specify it in a query algebra. While there are dozens of systems that implement one of the above ideas, none has seen widespread adoption. One reason may be the complexity of these systems.

For example, time series algebra systems require the user to express their query in formal language that may not be intuitive to a user who an expert in their domain, but may have little knowledge of SQL or other formal query languages.

In addition, the systems may simply not be *expressive* enough to be useful. For example, almost all query-by-sketching/example systems use some distance measure, typically Dynamic Time Warping (DTW) or a variant thereof as the underlying search mechanism [39]. However, there are simple intuitive patterns that can be easily specified in informal English, but not with any known query-by-sketching/example or time series query algebra paradigm. Figure 1 shows such an example in an entomological dataset.

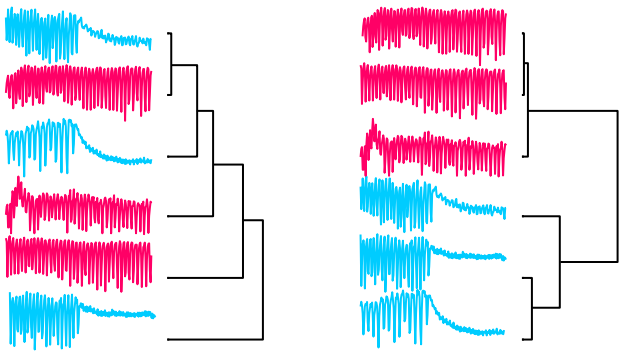


Figure 1: Six snippets from a dataset of behavior of a Beet Leafhopper (*Circulifer tenellus*), an insect pest for many commercially important crops. The behavior shown in cyan is known as Non-Ingestion-D1. *left*) If we attempt to cluster these behaviors with Euclidean distance the result is effectively random. *right*) A clustering based on natural language features is more intuitive (details later in this work).

If we use any of the cyan patterns as our query, query-by-content will *not* return a semantically similar pattern. However in the natural language space, the term `periodic followed-by constant` produces correct results.

The rest of this paper is organized as follows. In the next section we take the time to develop the reason *why* natural language search is such a promising technology for time series. In Section 3 we introduce the necessary notation and definitions and we introduce our search system in Section 4. We highlight the utility of our system in Section 5 by showing how we exploit the classic text retrieval idea of relevance feedback. In Section 6 we perform an extensive empirical evaluation. We will revisit related work in Section 7. Finally, we summarize our results and outline the future work in Section 8.

2 Why Natural Language for Time Series?

Given that there are dozens of query-by-*sketching* (*QbS*) [39][37], and hundreds of query-by-*example* (*QbE*) [31] systems for time series [13], it is natural to ask why we might also need a natural language query mechanism.

Consider the analogues that Google offers for image search. Google offers both natural language image search and query-by-example (*reverse image search* [40]). In Figure 2 we show the results of a query "`barefoot girl riding bike`".

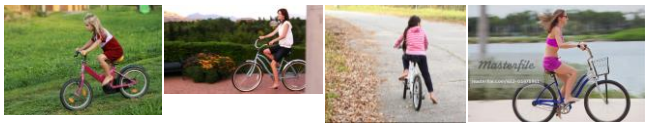


Figure 2: The top-four results for a Google image search of "*barefoot girl riding bike*".

By any standards, this is an extraordinary successful result. Of the (conservatively) trillions of images that Google has indexed, a vanishingly small fraction of them feature barefoot girls riding bicycles. Nevertheless, this query achieved almost perfect precision (upon careful inspection of the video from which the second image in Figure 2 was culled, the young lady was wearing a pair of incredibly gossamer sandals).

In contrast, Google's query-by-example is much less impressive. One can upload a photograph as a query and find exact *bit-for-bit* matches. In addition, because its main use is finding violations of copyright, it is also somewhat invariant to the obvious obfuscations someone might do to frustrate such a search, adding/removing a watermark, cropping or slightly changing the aspect ratio or color balance, or as shown in Figure 3.*left*, reversing the image.[1]

Figure 3: *left*) The top result of a reverse search of the rightmost image shown in Figure 2. Note that the images differ by the addition of a watermark, and by being reversed. *right*) A screen capture from [36] showing a query-by-sketching system.

While Google image search is able to find slightly corrupted version of images, it cannot find *semantically* similar images. For example, the last image shown in Figure 2 would never retrieve the *other* images shown in Figure 2, even though they are clearly related.

One could imagine a more sophisticated query-by-sketching system. Figure 3.*right* shows one of the most highly cited examples [36]. However, it is not clear that this system would retrieve either the second image in Figure 2 (occluded wheel) or the third image in Figure 2 (the bike is viewed from the rear). Moreover, while most people could draw a bicycle in its canonical side view, clearly drawing and indexing the concept "`barefoot girl`" is beyond the intended scope of [36], and beyond the artistic ability of most people.

There has recently been an explosion of interest in natural language querying of relational databases. We refer to such work as *QbV* (Query-by-Voice) [14][38][42]. Our proposed work shares a little of the motivation of *QbV*. We are not particularly motivated by the possibility of hands-free access [14] or providing access for individuals with disabilities (while recognizing that as a laudable goal). Our core claim is that in many circumstances, natural language time series search is more expressive than any *QbE* or *QbS* system.

[1] Here the archivist did not seem to realize that if one flips a bicycle image, the chain and chainring appear on the wrong side of the bike.

For example, if a user wants to find a noisy subsequence, she cannot simply draw a noisy query. Figure 1 shows that even if the user carefully selects one of the D1 behaviors as an example, she should not expect it to retrieve the other examples of that behavior. Likewise, no *QbE* or *QbS* systems that we are aware of can retrieve patterns based on a combination of local and global features, for example, retrieving a pattern that is high (relative to the *global* data), and has falling trend (a *local* feature).

3 Definitions and Notation

We begin by describing the necessary definitions and notation. The data type of interest is *time series*:

Definition 1 (Time series): A time series T of length n is a sequence of real-valued numbers t_i: $T = t_1, t_2, \ldots, t_n$.

Time series are often *multidimensional*:

Definition 2 (MTS): A multidimensional time series **MTS** consists of k time series T of length n, where $k \geq 2$.

MTS = $\{T_1, T_2, T_3, \ldots, T_k\}$,
where $T_1 = t_{11}, t_{12}, t_{13}, \ldots, t_{1n}$
$T_2 = t_{21}, t_{22}, t_{23}, \ldots, t_{2n}$
...
$T_k = t_{k1}, t_{k2}, t_{k3}, \ldots, t_{kn}$

t_{ji} and t_{ki} are two points cooccurring and $j \neq k$.

In many cases, the **MTS** and all the constituent time series have mnemonic names, for example a Body Area Network might be:

BAN = {*ECG*, *Temperature*, *Glucose*, O_2*Saturation*}

Where appropriate, we will use such mnemonic names in our examples. In many types of data analytics, in both the single and multidimensional cases, we do not care about *global* properties of the time series, but instead we are interested in the behavior of *local* regions. A local region of time series is called a *subsequence*:

Definition 3 (Subsequence): A subsequence $T_{i,m}$ of a time series T is a continuous subset of the values from T of length m starting from position i. $T_{i,m} = t_i, t_{i+1}, \ldots, t_{i+m-1}$, where $1 \leq i \leq n - m + 1$.

Our basic plan is to define a set of intuitive features that map each subsequence to a value. As shown in Figure 4, we do this for every subsequence, producing a *word feature vector* or "meta-time series" that describes each time series.

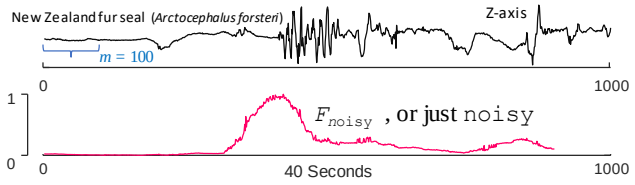


Figure 4: *top*) A snippet of seal behavior, captured with an accelerometer. *bottom*) A word feature vector $F_{\texttt{noisy}}$ corresponding to `noisy` that annotates the raw data. Note that the value of $F_{\texttt{noisy}}$ rises as it encounters the noise in the raw time series and falls as it encounters smoother data.

We call such time series *word feature vectors*:

Definition 4 (Word feature vector): Given a time series T and a function X, a *word feature vector* F is meta-time series of

length $n - m + 1$ that annotates T, reporting the value of X for every subsequence of T.

Each of these word feature vectors corresponds to an intuitive property of a time series. For example, we have $F_{\texttt{noisy}}$, $F_{\texttt{smooth}}$, $F_{\texttt{linear}}$ etc. To allow a more intuitive use of our system, we will use the English terms such as noisy, smooth, linear etc to refer to the vector, as in `noisy`, `smooth`, `linear` etc.

We need to review one issue that all time series subsequence ranking systems must deal with [29][33], *trivial matches*:

Definition 5 (Trivial matches): Given a time series T, containing a subsequence $T_{p,m}$, if $T_{p,m}$ scores highly on any scoring function, then $T_{j,m}$ which $j \epsilon [-m/2, m/2]$ will almost certainly score high on the same function. These spurious high scoring subsequences are *trivial matches*.

To avoid counting trivial matches when finding the top-k matches to a query, we discard some of the patterns using the concept of an *exclusion zone,* a standard practice [29][33].

We list and briefly define below a subset of initial set of features we have implemented. Note that in every case the feature vectors are normalized to be between zero and one. Where features come in natural pairs, we define only one. Unless otherwise stated, the negative feature X can be obtained by simply computing $X = 1 - Y$, where Y is the positive version of the feature.

The formal definitions and code for each feature are relegated to [16] for brevity. We begin with the *local* features:

- **Local Features**
  - `rising` (`falling`): The degree to which the slope of the best fit line is positive [10].
  - `concave` (`convex`): The degree to which there is a low residual error when modeled by 2nd degree polynomial.
  - `linear` (`non-linear`): The degree to which there is a low residual error when modeled by 1st degree polynomial.
  - `smooth` (`noisy`): The degree to which there is a low residual error when modeled by a moving average.
  - `complex` (`simple`): Complexity [3] rewards a subsequence for having more peaks, valleys and features.
  - `spiky` (`dropout`): Rewards a subsequence for having a small (but non-zero) number of data points much greater than (less than) the mean.
  - `periodic` (`aperiodic`): The degree to which there is a low residual error when modeled by two FFT coefficients[13][21].
  - `symmetric` (`asymmetric`): Rewards a subsequence for having an axis symmetry about its midpoint.
  - `step` (`no-step`): Rewards a subsequence for being modeled well by a small number of segments using a Piecewise Constant Approximation [4].
  - `high-amplitude` (`low-amplitude`): Rewards a subsequence for having a high variance.

Figure 5 shows some examples of subsequences from various real datasets that have a high score on one of our features. Note that some of the features can come in two "flavors", normalized and unnormalized [15][29].

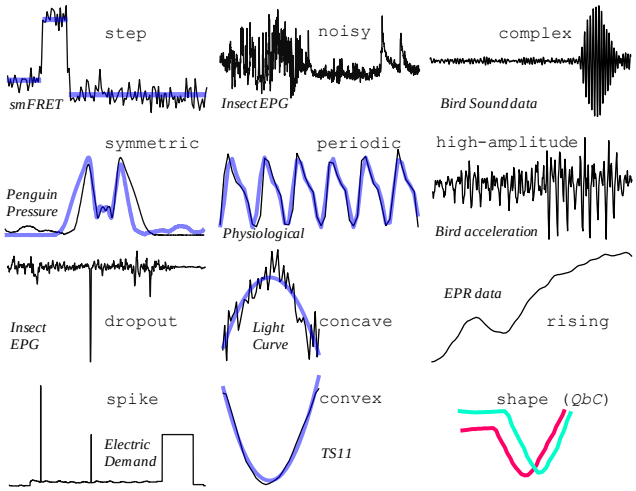


Figure 5: Subsequences from various datasets that illustrate one of our features. Note that in general, scoring high on one feature does not preclude scoring high on others. For example, *Penguin-pressure* would also score high on `smooth`.

The local features can be computed by looking only at the relevant subsequence. However, there are also some features that can only be computed relative to a global context:

- **Global Features**
  - `high` (`low`): Rewards a subsequence for having a relatively high value compared to the overall time series.
  - `typical` (`unusual`): Rewards a subsequence for being unusual [6] (a low value in the Matrix Profile [29]).

One issue we may face is the lack of agreement on the correct wording for the features. These differences may be idiosyncratic, or domain informed. For example, the oil and gas community often refers to `low-amplitude` regions as *flat* or *level*. We can address this with a synonym list. Thus we have:

```
low-amplitude: flat, level, constant,...
```

Our thesaurus is currently handcrafted, but there is rich literature in learning thesauruses automatically from the web or from user interactions [20]. Finally, as illustrated in Figure 6 and Figure 7 we support two *special* features:

- **Special Features**
  - `keyword`: For time series data which are annotated by text we reward a subsequence for being covered by that keyword. For example, in medical datasets, `keyword`("sleeping") would reward a subsequence that falls in a region covered by a nurse's free text annotation "*patient is sleeping, skin pale.*"
  - `shape`: For specialized domains we can define a dictionary of shapes and reward a subsequence for resembling them under any given distance measure. For example, `shape`(ValveFailure) could reward a subsequence which resembles a shape that a user associates with a previous encountered problem [27].

In Figure 6 we show an example of a word feature vector created using a keyword. Here we again support the use of synonyms. For example, in a medical domain we set the value of

$F_{\texttt{keyword}}$("jaundice") to be "1" if we encounter "jaundice" or "yellow" or "icterus".

In medical domains, specialized thesauruses have existed for over one hundred years [28], and can now be accessed with APIs. For more general domains, we may have to learn or construct an appropriate mapping [5].

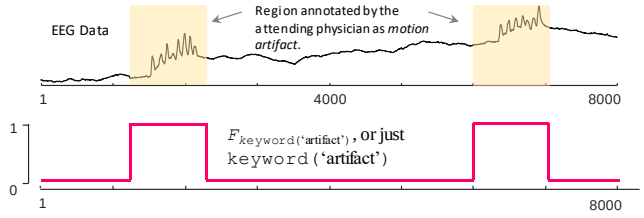


Figure 6: *top*) A snippet of EEG data [24] that was hand annotated by a physician. *bottom*) The word feature vector that was created to reflect the doctor's annotation.

Our final feature is `shape`. Inclusion of this feature allows our system to subsume most existing *QbS* and *QbE* systems for time series [37][31][39]. Figure 7 illustrates the basic idea.

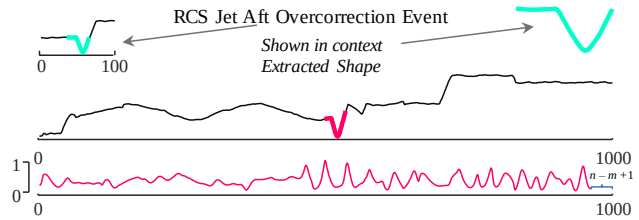


Figure 7: *top*) Extracting a specific shape from the STS-57 space mission time series. *middle*) The time series T with just the best matching location highlighted. *bottom*) Computing the shapes distance to all subsequences in T allows us to compute the `shape`("overcorrection") feature which annotates all the subsequences as to how similar they are to the target pattern.

Here an engineer found an "overcorrection" pattern in one trace that she believes is important enough to be a named primitive in future searches. We can use the MASS algorithm efficiently compute its *distance profile* [15] relative to T, which is simply its z-normalized distance to every subsequence in T. Because we use high values to indicate the presence of a feature, we "flip" the distance profile before normalizing it.

In summary we call these word features a *feature-set*. We also provide a dictionary of synonyms for each feature within the feature-set called *dictionary-set*. While we precompute and store these word features, the MASS algorithm to compute the distance profile is so efficient (under a second for $n = 2^{22}$) that we could also support ad-hoc queries[37][31].

3.1 The Temporal Operator

In addition to the features mentioned in Section 3, we allow a temporal operator `followed-by`. This operator can be parameterized to include the *maximum* number of data points the user expects feature *B* follows feature *A*. Suppose we have a time series sampled at 60 Hz. If the user is looking for the highest increase in value that happens in under two seconds, she could query `low followed-by(120)high`. With no parameter given, it is defaulted to the subsequence length.

Our single operator may seem limited relative to the expressiveness of Allen's interval algebra [1]. However, consider the following: Google search supports forty-eight advanced search operators. However, even the three simplest and most intuitive Boolean operators are very rarely used.[2] People expect to be able to answer almost any query with just an (implicit) AND operator.

Moreover, in working with experts in domains as diverse as entomology and petrochemical production, we were struck at how universal and natural the `followed-by` descriptor was. When we asked "*How would you recognize an event of type X in your data?* ", almost everyone said something like, "*You will see an A, followed by a little bit of B, followed by a lot of C* ."
The synonyms for our temporal operator are:

```
followed-by: succeeded-by, then, next, ...
```

The taxonomy above is by no means meant to be complete. The reader may have ideas for additional features which could be added to our framework. However, as we will show, this vocabulary is expressive enough to allow us to find targeted patterns in dozens of diverse domains.

3.2 A Worked Example of Defining a Feature

The formal definitions/code for each feature are relegated to [16]. We show some of the introspection and effort that went into defining one feature, `spiky`. Even for a simple feature such as `spiky`, "*the best method to identify spikes in time series is not known.*" [7].

Our first attempt to define the `spiky` feature used the intuition that a spike is an abnormally high value. This can be captured by smoothing the time series, then calculating the residual error between the original and smoothed versions. The high residual error indicates an abnormally high value in the time series. This can be obtained by:

```
ResidualError = T - smoothing(T);
spiky = sum(ResidualError);  % Version 1
```

There are several smoothing functions that can be used for this purpose but after testing several smoothing functions we chose `medfilt1`, the built-in function in Matlab.

We realized this definition has several weaknesses. Sequences that most people would judge as noisy but not spiky can achieve a high score. This is because summing the ResidualError in one noisy subsequence can have the same effect as the spiky subsequence. The simple solution to this problem is taking the *maximum* of ResidualError instead of summing them. In Matlab this can be captured by:

```
spiky = max(ResidualError); % Version 2
```

Empirically, we find that this definition is much more robust. Many of the features we implemented required multiple iterations to produce acceptable results. We do not claim that any of our definitions are objectively *correct*. As [7] points out, even for something as simple as *spike*, there may be different definitions that are best for epidemiologists vs. economists etc.

[2] For Google search, AND is implicit, OR needs to be explicitly added, and NOT is represent as -.

4 NLP Time Series Search

Recall that the *feature-set* is a set of local and global word features, and the *dictionary-set* are key-value pair synonyms for each feature within the *feature-set* as shown in Figure 8. For each feature in the *feature-set* there is a corresponding function which takes a multidimensional time series *MTS* and the subsequence length *sub* as inputs and returns a meta-time series *index-feature* (e.g. $index_{\texttt{rising}}$).

We are now in a position to explain our system. At a high-level, for each time series, we precompute an *index*, which is an array containing all of the *index-features*, as shown in Figure 8. This *index* can be stored on a disk to await future query sessions. At query time, we begin by *parsing* the query to ensure that it is in the correct format. Then we use a search algorithm to find the top-k subsequences that are best matches for the query. Below we explain these steps in more detail.

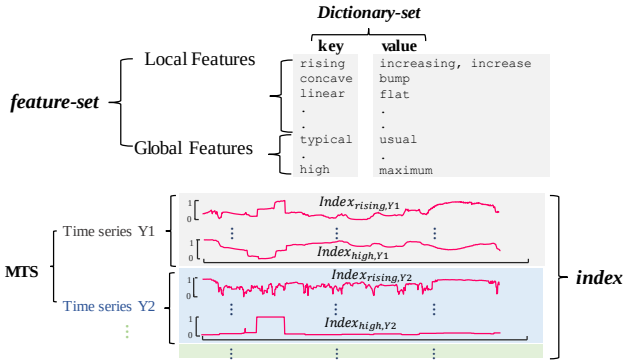


Figure 8: *top*) An illustration of the *feature-set* and *dictionary-set*. *bottom*) The *index* for the given time series Y1, Y2 etc.

The inputs to the algorithm *Build-Index* are a multidimensional time series *MTS* of size p × q and a user-defined subsequence length *sub*. The output of the algorithm is the *index*, which is a p × (q − *sub*) array normalized between zero and one. The algorithm *Build-Index* is outlined in Table 1 and its output is hinted at in Figure 8.

In line 1, the algorithm initializes an *index* array *Idx* to an empty array. In lines 2 to 5 we calculate the *index-feature* for each feature in the *feature-set*, then add it to the *index* array *Idx*. The *index* array *Idx* is returned on line 6.

Table 1: The *Build-Index* Algorithm

Algorithm: *Build-Index* (*MTS*, *sub*) **Input**: Multidimensional time series *MTS*, subsequence Length *sub* **Output**: Index *Idx*
1: *Idx* ⟵ empty array 2: **for** *feature* **in** *feature-set* **do** 3: *Index-feature* ⟵ feature(*MTS*, *sub*) 4: *Idx* ⟵ *Index-feature* 5: **end** 6: **return** *Idx*

After building the index we store it on the disk (discussed in Section 4.2). Now we need a valid query to search the index. However, user input can be imperfect, containing spurious punctuation or misspelled words. Thus, we need to parse the text to produce a valid query. For the ease of explanation, we initially ignore the `followed-by` term in the query.

At query time we break a sequence of characters into pieces called *tokens*. In order to tokenize a multidimensional time series the user needs to specify *which* dimension(s) she is interested in. We resolve this by asking the user to specify the name of the time series and features of interest all in uppercase and lowercase letters, respectively. For example, for the query: `TEMPERATURE high spike PRESSURE linear`, we parse it into six tokens, with the uppercase letters `TEMPERATURE` and `PRESSURE` indicating the time series name and the features [`high`, `spike`] and [`linear`] being the corresponding features for each time series respectively.

For any token, if we cannot find the corresponding feature we search the *dictionary-set* to find the key indicating its synonym as shown in Figure 8. If we cannot find the token in the *dictionary-set* we simply ignore it but echo a warning to the user i.e. "*missing feature* `bpike`". Now we are ready to find the top-k results.

In the spirit of information retrieval from which we draw inspiration [2], each subsequence can be considered a *document*. After parsing, for each document j we calculate the score^j, similar to text mining, as the sum of all the *index-features* within the query.

$$\text{score}^j = \sum_{i \in \text{Query}} Index_i{}^j$$

The algorithm *Top-K-Results* for finding the best matches to a query is outlined in Table 2. The inputs include a query Q and a user-chosen number of k matches. The output is the top-k subsequence within the time series of length *sub*.

We initialize an array *score* to an empty array in line 1. We iterate over each feature in the query Q and add the *index* for that specific feature to the *score* array in lines 2 to 4. Then we store the total score in s which is the element wise summation of the *score*. In lines 6 to 9 we sort s in the descending order and find the k subsequences, *top-k*, within time series as we are taking care of the exclusion zone. Finally the *top-k* results (i.e. k subsequences) are returned.

We can now explain how we handle the `followed-by` case. If the query includes any $\texttt{followed-by}_{\texttt{value}}$ terms, we truncate the *index-feature* by the desired parameter amount and shift the *index-feature* relative to all other *index-features* in the *index*. If the user does not provide a value, we default to the subsequence length *sub*.

$$\texttt{followed-by}_{\texttt{value}}\ Index = Index(\text{value: end} - \text{value} + 1)$$

In order to keep the same size *index-feature* as the original one, we pad zeros to the end of the just calculated *Index-feature*. We then give this modified *index* as an input to Table 2 to calculate the top-k best matches.

Table 2: An Algorithm to Find the Top-K-Results

Algorithm 2: *Top-K-results* (*Q*, *k*) **Input**: The user defined query *Q*, number of results *k*, build-Index *index*, subsequence length *sub*, Multidimensional time series *MTS*

Output: Top-k subsequence *top-k*

1: *score* ← Ø, *top-k* ← Ø
2: **for** *q* in *Q* **do**
3: *score* ← *index*(q)
4: **end**
5: *s* ← sum(*score*) // element-wise summation of *score*
6: *id* ← sort (*s*, '*desc*end')
7: **for** *i* in *k* **do**
8: *top-k* ← *MTS* [id_i : id_i+*sub*] //subsequence of length *sub*
9: **end**
10: **return** *top-k*

4.1 Cold Starting our System

Our system is simply a *text* search, and 54% of the US population performs several text searches each day [22]. Nevertheless a novice user may not have the vocabulary to create an effective query. To solve this cold start problem, we can switch our system into *suggest* mode, in which the user can use her pointer to indicate any location in a time series. Then, the tool instantly displays a sorted list of the feature words that would have (individually) scored the relevant subsequence highly [16]. After a short amount of time interacting with this system, the user can gain intuition as to what *compound* queries may satisfy her query needs.

4.2 Indexing for NLP Time Series Search

We have optimized our system for simplicity and scalability. We claim that we do not need to create sophisticated indexes for our representation (for most users). Recall our insect example, recorded at 100 Hz. The longest example is eleven hours long, or 3,960,000 data points. Suppose we extracted all our feature vectors and stored them on disk, and at query time, a user queries "`concave spiky`". Including the time needed to load the data from the disk, we can answer this query in under one second using nothing more than the optimized brute force search outlined in Table 2.

However, the *memory* overhead could become a bottleneck. The space overhead is $O(nk)$, which for longer time series becomes untenable. There is one piece of low hanging fruit that we can exploit. Recall that most[3] of our features come in complementary pairs. That is to say featureA = 1 – featureB. Thus, we can store only one of the pair, and when needed, invert its value, effectively halving our memory requirement.

To achieve further compression of the feature vectors we can exploit the rich literature on time series dimensionality reduction [30][23][26]. Most time series are amiable to a significant dimensionality reduction with little error. Many techniques have been proposed, including Discrete Fourier transform (DFT) [30], the Discrete Wavelet Transform (DWT) [23], Piecewise Aggregate Approximation (PAA) etc. Figure 9 show an example of a feature vector approximated with DFT.

[3] *Most*, but not *all*, `spiky` ≠ (1 - `dropout`)

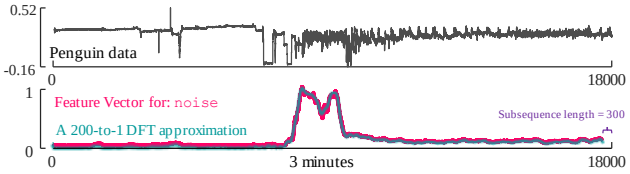


Figure 9: *top*) A 3-minute snippet of time series. *bottom*) The `Noise` feature vector for the penguin data (red), and its DFT 200-to-1 approximation (blue)

While different representations can better suit different kinds of data [23], empirical comparisons suggest that on average, there is no reason to favor any approach in terms of reconstruction error [26]. However, different representations have other properties that can be useful depending on how they are deployed. For our system we use the DFT. The DFT representation has the property that it can be very quickly cast back from its reduced representation to the original dimensionality. To concretely ground this with some real numbers, consider the following. Recall our eleven hours/3,960,000 data points of insect data. We compute one feature vector for it, we compress it at a twenty-to-one ratio, and then we save it to a disk. If we are given a query that requires this feature vector, we can retrieve it from a disk, and upsample it (i.e., perform an inverse FFT) in 0.12 seconds. So even if our natural language query involved five keywords, and all data is disk-resident, we can comfortably return the answer in under a second.

Since the number of keywords we currently support is about thirty, and our complimentary observation reduces the number of vectors we must save to about half that, a 15-to-1 compression gives us a space overheard of just $O(n)$.

To test the scalability of this scheme, we performed the following experiment. We created, then saved to a disk, all the feature vectors for an increasing longtime series. We used EOG data concatenated from different individuals, because it is a highly variable data source, which typically features a "bit of everything".

We created queries by randomly selecting four words from our vocabulary, for example {`concave`, `spiky`, `rising`, `symmetric`}, and measured how long it took to find the single best match. As a point of comparison, we also measured the time needed to perform a one-nearest Euclidean distance neighbor query on the original data, assuming all data is in main memory. For this we used MASS, which is the optimally fast algorithm[4] [15]. We used random queries of length 500, although MASS is essentially invariant to the query length and the data's structure. Figure 10 shows the results.

[4] This claim needs two qualifications. If we know the query length ahead of time (and its intrinsic dimensionality is small), and we are given a large memory budget, we could build an index that might support faster nearest neighbor queries. In the more general case, for ad-hoc queries whose length is not known ahead of time, MASS *is* optimal if (as widely believed) FFT is optimal [15].

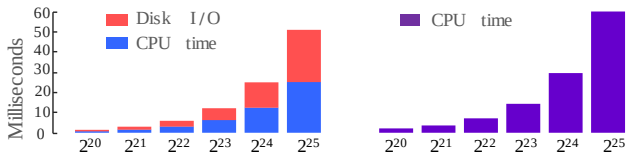


Figure 10: *left*) The time taken to answer natural language queries, with a disk resident index, for increasing long time series. *right*) For the same time series, the time taken to answer a one-nearest Euclidean distance neighbor query on the original data, using the optimal algorithm [15], and assuming that disk access take zero time.

These results show that even if we use a level of dimensionality reduction to bring our space complexity down to $O(n)$, we can answer queries for very large datasets in an *interactive* time. Moreover, the time needed to answer queries compares very favorably with the state-of-art query-by-content methods [15].

Given that we can support interactive time queries for datasets with millions of data points, our incentive to further optimize is diminished, as we have a tool that can support the entomologists, medical doctors, marine biologists etc. that motivated this work. Clearly there are potential "customers" for natural language systems that may desire scalability to hundreds of millions of data points and beyond. For data sizes just beyond our current grasp, the low-hanging fruit of parallel computation and/or GPUs offer the possibility for an easy order of magnitude improvement. Beyond that, novel representations and data structures will be needed, perhaps augmenting and adapting work on indexing intervals [11]. We leave such considerations to future work.

5 User Feedback: A worked example

User feedback can be incorporated in text search to improve the percentage of the relevant document displayed to the user [2]. One feature of our system is that it is amiable to most text retrieval techniques, including relevance feedback [2][18].

Consider the time series in Figure 11, which represents the behaviors *walking*, *running*, and *skipping* from PAMAP datasets.

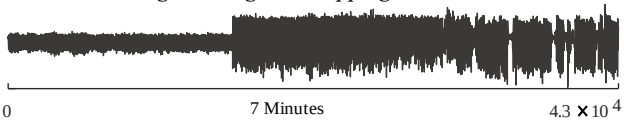


Figure 11: The hand-mounted X-axis acceleration of an individual walking, running, and then skipping rope [34].

Imagine our user wants to find a particular pattern that she believes will discover *walking* behaviors, and she initially searches with: `X-ACCELERATION, symmetric, concave`, examining just the first top-6 results, as shown in Figure 12.

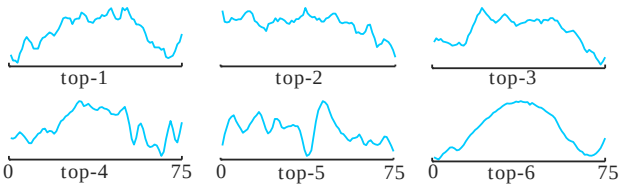


Figure 12: The top-6 results of the user query, `X-ACCELERATION, symmetric, concave`, on the dataset shown in Figure 11.

In this case, the results are something of a mixed bag. Can we improve them? One way to achieve this is to use Rocchio's Algorithm [18], exploiting our text-based view of time series [19]. This algorithm revises the original query based on the feedback that the user gives about the relevance of the returned results. The modified query moves closer to the centroid of relevant documents. The algorithm has proven convergence properties, given only the assumption that the relevant and non-relevant documents form clusters and the notion of their centroid is meaningful is certain senses [18][19]. However, empirically it often works even if those assumptions are violated. The formula for Rocchio's Algorithm is:

$$\overrightarrow{Q_J} = (a.\overrightarrow{Q_0}) + \left(b.\frac{1}{|D_r|}.\sum_{\overrightarrow{D_j}\epsilon D_r}\overrightarrow{D_j}\right) - \left(c.\frac{1}{|D_{nr}|}.\sum_{\overrightarrow{D_k}\epsilon D_{nr}}\overrightarrow{D_k}\right)$$

$\overrightarrow{Q_J}$ is the modified query, $\overrightarrow{Q_0}$ is the original query and $|D_r|$ and $|D_{nr}|$ are the size of relevance and non-relevance documents, respectively. $\overrightarrow{D_J}$ and $\overrightarrow{D_k}$ are sets of vectors containing the coordinates of relevance and non-relevance documents. a, b, and c are the hyper-parameters. For instance, setting "a" to zero will exclude the original query from the next result. Assume our user chooses just result number 6 as a relevant document. The result of the modified query is shown in Figure 13, with $a = 1$, $b = 0.8$, and $c = 0$, typically recommend values in the literature [18][19].

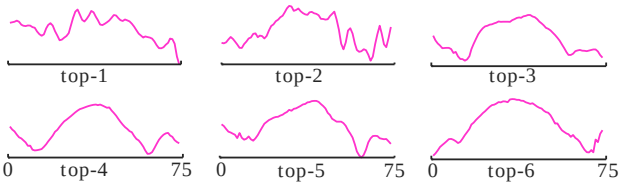


Figure 13: The top-6 results using Rocchio algorithm.

Beyond the subjective visual superiority of these patterns, by checking with the contemporaneous annotations provided with this data, we discover that 4 of 6 results now *correctly* reflect the query purpose i.e. *walking* behavior.

As impressive as this result is, we can improve it even further by using Euclidean distance in our measurement. Recall that an interesting feature of our system is that we subsume most query-by-sketching [39][37], and query-by-example [31] systems by incorporating the `shape` feature.

In text mining, the Cosine similarity is commonly used to measure the similarity between two documents. An important property of Cosine similarity its independence from a document's length. In our system we can compare subsequences

in the same manner as a comparison between documents in text mining. The subsequences are the same length which allows us to use Euclidean distance as a similarity measure [3]. In this case we revise the top-k results by computing the Euclidean distance between a word feature vector of each subsequence and a word feature vector of the centroid of the relevant documents. We modify the result as follows:

$$S^k = (a. \sum_{w_i \in Q_0} FV_{w_i}^k) - b. \sqrt{\sum_w (FV_w^k - FV_C)^2}$$

k is the subsequence index and S^k is the score for the k^{th} subsequence. $\sum_{w_i \in Q_0} FV_{w_i}^k$ is the sum of feature vector for each word within the original query in subsequence k.

We calculate Euclidean distance between the feature vector for each word in our dictionary, FV_w^k, and the average feature vector for the relevant documents, FV_C . a and b are the hyper-parameters for modifying the score. To produce the top-k result we sort S^ks in descending order, producing the top-6 result shown in Figure 14.

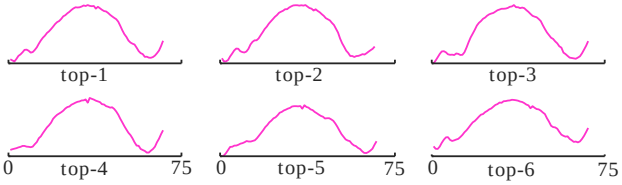


Figure 14: Top-6 results of user defined query using Euclidean distance as a similarity measure.

Not only are the results subjectively improved, but a post-hoc analyses confirms that now *all* these subsequences belongs to the same *walking* behavior.

6 Experimental Evaluation

To ensure that our experiments are reproducible, we have built a website which contains all data/code/raw spreadsheets for the results, in addition to many additional experiments that are omitted here for brevity [16]. This commitment to reproducibility extends to all the examples the previous sections. Because the ability to search time series with natural language is new, we first conduct short case studies to help the reader appreciate the utility of this ability. Except where otherwise stated, in the below examples, we set the top-k to reflect the number of true positives in the dataset, to make the numerical evaluation of accuracy easier.

6.1 Marine Mammals

We consider a seal behavior dataset [12] which contains twenty-six distinct behaviors of seventy-two seals, belonging to four species. As Figure 15 shows, this dataset contains data from a wearable accelerometer mounted on the seals back.

We must immediately disclaim, none of the authors has any experience or knowledge in this domain. However, as we shall show, *in spite of* our naivety, our system does allow us to ask and answer useful queries. To give the reader some appreciation of the structure of the data, Figure 15 shows the Y-acceleration Australian fur seal AFS (*Arctocephalus pusillus*).

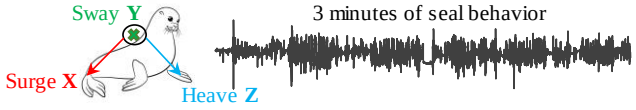


Figure 15: A sample of seal behavior hints at the complex structure of the data.

One of the behaviors observed by the seal's handlers is *scratching*. Scratching in seal species is described as "*perched on fore-flippers with rear on the ground, head down towards rear and one rear flipper scratching neck or head*" [12]. As non-experts, it is hard for us to translate this into a query, but we can bypass this issue with the following simple idea.

We watched videos to see examples of this behavior [41] and try to emulate it by "puppetry", while holding a phone to record the motion (Matlab has an iPhone app that allows you to capture all the phone's sensors as time series).

After watching some videos of seals *scratching*, our first attempt at puppetry produced the pattern (~3.5 seconds). We decided to covert this to the text query `SWAY periodic` and `low-amplitude`.

The result for this query is shown in Figure 16. The green color is a binary vector encoding the ground truth for the scratching behavior appearing in the time series. The red regions within time series is the result of the scratching query.

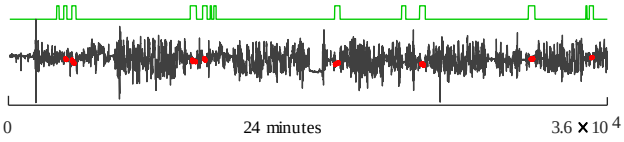


Figure 16: *top*) A Boolean vector representing the scratching behavior (green bar). *bottom*) The time series annotated (red) with the top-13 *scratching* query results.

Visually the results are very promising. We can objectively score them as follows. We call each separated labeled behavior a *block*. *A success-block* is a block that our algorithm correctly identifies as having the behavior. The accuracy is the sum of all the success-block length divided by the total blocks.

$$\text{Accuracy} = \frac{\sum_{j \in success-block} length\ of\ labeled\ behavior_j}{\sum_{i \in block} length\ of\ labeled\ behavior_i}$$

The accuracy of our method is 78.3%. In contrast, the default rate (random sampling) is just 0.66%

Another behavior in the seal dataset is *shaking*. Shaking is described by the experts as a "*short and sharp movement of the head left and right to remove water from the fur*" [12]. For this behavior a seal sits on the ground and after a couple of seconds of sudden movement, the seal sits at rest again.

The shaking query can be transcribed as `SWAY constant followed-by complexity followed-by constant`. Figure 17 shows a search for this within a concatenation of Y-acceleration from eight New Zealand fur seals (*A. forsteri*).

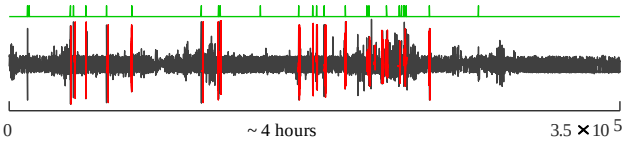


Figure 17: *top*) A Boolean vector representing the *shaking* (green). *bottom*) A time series annotated with red regions are the result for the top-26 *shaking* query.

The *default rate* for finding shaking behavior is just 0.01%, but the accuracy of our proposed method is 80.7%. As this behavior was perhaps the most amenable to shape-based search, we also compared the Euclidean distance similarity measure to our result (i.e. query-by-*example* [31]), however Euclidean distance achieved just 37.0%.

The final behavior we consider for the seal dataset is a terrestrial gait called *moving*. Moving is described as "*locomotion out of water which four flippers used as legs*". Here we wish to test the following hypothesis: Given a concrete example of a behavior in one animal, can we construct a query that will generalize to a different member of the same species? We examined the annotated time series of Australian sea lion ASL (*Neophoca cinerea*) to extract this pattern . Using this pattern, we describe the Moving behavior as `SWAY constant followed-by rising followed-by falling followed-by constant`. Figure 18 shows the result of this query on a *different* seal.

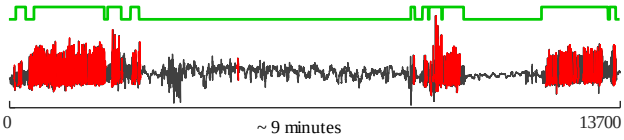


Figure 18: *top*) A Boolean vector representing the moving behavior (green bar). *bottom*) Time series of Y-acceleration of Abbey (sea lion) annotated with red regions are the result for the top-230 *movement* query.

While the previous experiments use *accuracy* as a metric, here we use the *success-rate*. In the previous experiments each behavior happens once or twice per block. For this experiment each block contains a lot of moving behavior, so observing just one does not explain the whole block. We define success-rate as a number of times we correctly predict a query within the green region divided by the total number of trials. The success-rate for our system is 95.6%. Visually we can see that most of our results are true positives. We also compare our result with two similarity measures: Euclidean distance *ED* and *DTW* [39]. The success-rate of these methods is 20.1% and 30.0%, respectively. Figure 19 visually summarizes this.

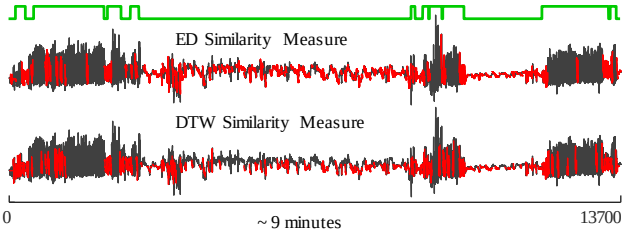


Figure 19: *top*) A Boolean vector representing the moving behavior (green bar). *bottom*) The time series shown in Figure 18. The red subsequences show the top-230 results for ED and DTW are plagued by false positives.

7 Related Work

In [17] the authors introduce a shape description language (SDL) with a formal semantics for searching time series. Like our approach, it has an alphabet of primitives, including `up`, `Up`, `down`, `Down`, `Stable` etc. However, each primitive is Boolean. That is to say, each subsequence either has the `Up` property or it does not. In contrast, our approach allows *continuous* degree of membership to `Up` (we call "`rising`"), which we argue is more robust and expressive. For example, a query to SDL can return the empty set, whereas we rank every possible subsequence in the dataset.

Beyond the brittleness of SDL, the language itself may be too precise and complex to be useful in many contexts. For example, to find a region with two peaks, our system uses `peak followed-by peak`, whereas SDL uses [23]:

```
Shape doublespeak (width ht1 ht2) (in width (in
order spike (ht1 ht1) spike (ht2 ht2))))
```

The original SDL paper did not test on real data, and while it is highly cited, none of the citing papers directly implements these ideas. However, a recent paper does implement a similar system called ShapeSearch [25]. The ShapeSearch system has a similarly expressive query language, for example to find "*an overall upward trend between x values 1 to 10 with a peak between 3 to 5.*" The user can type:

```
[p{up},xr{1,10}] AND [p{peak},xr{3,5}]
```

Again, we see such precision as being very useful in certain limited domains, but too complex and unforgiving for many.

There are many other interesting (*QbS*) [39][37], and (*QbE*) [31] systems for time series [13], but to our knowledge there are none that directly compete with our system.

8 Discussion and Conclusions

We have introduced a natural language search system for time series data. Our system is intuitive and easy to use, yet arguably more expressive than any existing query-by-*sketching* or query-by-*example* system [37][31][39]. Future work includes *learning* the mappings between the data and the words [5][32] and summarizing the time series [8] using words.

REFERENCES

[1] J.F. Allen (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 832-843.

[2] R. Baeza-Yates. and B. Ribeiro-Neto, 1999. *Modern information retrieval* (Vol. 463). ACM Press, New York, NY.

[3] G.E. Batista, E.J. Keogh, O.M. Tataw, and V.M. De Souza, (2014). CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28 (3), 634-669.

[4] K. Chakrabarti,, E. Keogh,, S. Mehrotra, and M. Pazzani, (2002). Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems (TODS)*, 27(2), 188-228.

[5] M. De Rijke. 2017 Apr 3-7. Learning to Search for Datasets. *26th International Conference on World Wide Web*. Perth, Australia.

[6] Anastasia Giachanou and Fabio Crestani. 2016. Tracking Sentiment by Time Series Analysis. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 1037-1040.

[7] D.E. Goin and J. Ahern. 2018. Identification of Spikes in Time Series. arXiv: preprint arXiv:1801.08061.

[8] S. Imani et al., "Matrix Profile XIII: Time Series Snippets: A New Primitive for Time Series Data Mining," in IEEE Int. Conf. on Data Mining (ICBK2018), 2018.

[9] Eamonn J. Keogh, Michael J. Pazzani. 1999. Relevance Feedback Retrieval of Time Series Data. SIGIR, 183-190.

[10] Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, and Zhe Zhao. 2014. Predicting bursts and popularity of hashtags in real-time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (SIGIR '14)*. ACM, New York, NY, USA, 927-930.

[11] O. Kostakis and P. Papapetrou. 2017. On searching and indexing sequences of temporal intervals. *Data mining and knowledge discovery*, 31(3), 809-850.

[12] M.A. Ladds, A.P. Thompson, D.J. Slip, D.P Hocking, R.G. and Harcourt. 2016. Seeing it all: evaluating supervised machine learning methods for the classification of diverse otariid behaviours. *PloS one*, 11(12), p.e 0166898.

[13] Zitao Liu, Yan Yan, and Milos Hauskrecht. 2018. A Flexible Forecasting Framework for Hierarchical Time Series with Seasonal Patterns: A Case Study of Web Traffic. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 889-892.

[14] G. Lyons, V. Tran, C. Binnig, U. Cetintemel, and T. Kraska. 2016, June. Making the case for query-by-voice with echoquery. In *Proceedings of the 2016 International Conference on Management of Data*. ACM, New York, NY, USA, 2129-2132..

[15] A. Mueen, K. Viswanathan, C.K. Gupta, and E. Keogh. 2015. The fastest similarity search algorithm for time series subsequences under Euclidean distance. url: www. cs. unm. edu/~ mueen/FastestSimilaritySearch.html

[16] Project Website: https://sites.google.com/site/nlptimeseries

[17] R.A.G. Psaila, Mohamed Wimmers and E.L. It. 1995. Querying shapes of histories. *Very Large Data Bases. Zurich, Switzerland: IEEE*.

[18] J.J. Rocchio. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*, 313-323.

[19] G. Salton, A. Wong, and C.S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

[20] S. Shekarpour, E. Marx, S. Auer,. and A.P. Sheth. 2017. RQUERY: Rewriting Natural Language Queries on Knowledge Graphs to Alleviate the Vocabulary Mismatch Problem. In *AAAI*, 3936-3943.

[21] Milad Shokouhi. 2011. Detecting seasonal queries by time-series analysis. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 1171-1172.

[22] Statista: The statistics Portal. Retrieved from https://www.statista.com/statistics/250948/frequency-of-google-search-usage-by-us-users/

[23] S. Sumathi,. and S.N. Sivanandam. 2006. *Introduction to data mining and its applications* (Vol. 29). Springer.

[24] K.T. Sweeney, S.F. McLoone and T.E. Ward. 2013. The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique. *IEEE transactions on biomedical engineering*, 60(1), 97-105.

[25] T. Siddiqui, P. Luh, Z. Wang , K.Karahalios, A. Parameswaran. 2018. ShapeSearch: Flexible Pattern-based Querying of Trend Line Visualizations (Demo). *VLDB'18: 44th Int'l Conf on Very Large Data Bases*, Rio De Janeiro, Brazil,.

[26] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh. 2013. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2), 275-309.

[27] Sheng Wang, Zhifeng Bao, J. Shane Culpepper, Zizhe Xie, Qizhi Liu, and Xiaolin Qin. 2018. Torch: A Search Engine for Trajectory Data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 535-544.

[28] M.B. Wilfred. 1903. A Thesaurus of Medical Words and Phrases. *The Laryngoscope*, 13(9), 736.

[29] C.C.M. Yeh, et. al. 2016. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets. *IEEE ICDM 2016*, 1317-1322.

[30] R. Agrawal, C. Faloutsos, and A. Swami. 1993, October. Efficient similarity search in sequence databases. In *International conference on foundations of data organization and algorithms*. Springer, Berlin, Heidelberg, 69-84.

[31] H. Hochheiser, and B. Shneiderman. 2002, April. A dynamic query interface for finding patterns in time series data. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM, 22-23.

[32] L. Soldaini, E. Yom-Tov. 2017 Apr 3-7. Inferring individual attributes from search engine queries and auxiliary information. In *26th International Conference on World Wide Web*. Perth, Australia, 293–301.

[33] B. Chiu, E. Keogh, and S. Lonardi. 2003, August. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 493-498.

[34] A. Reiss and D. Stricker. 2012, June. Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium*. 108-109.

[35] Haggai Roitman, Sivan Yogev, Yevgenia Tsimerman and Yardena Peres. 2013. Towards Discovery-Oriented Patient Similarity Search Health Search and Discovery (HSD) workshop (@SIGIR).

[36] C. Xiao, C. Wang, L. Zhang,. and L. Zhang. 2015, June. Sketch-based image retrieval via shape words. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval* (pp. 571-574). ACM, 571-574.

[37] M. Correll, and M. Gleicher. 2016, October. The semantics of sketch: Flexibility in visual query systems for time series data. In *Visual Analytics Science and Technology (VAST), 2016 IEEE Conference*. IEEE. 131-140.

[38] D. Chandarana, V. Shah, A. Kumar, and L. Saul. 2017, May. SpeakQL: Towards Speech-driven Multi-modal Querying. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM. 11.

[39] M. Mannino and A. Abouzied. 2018, May. Qetch: Time Series Querying with Expressive Sketches. In *Proceedings of the 2018 International Conference on Management of Data*. ACM. 1741 – 1744.

[40] Google Image Search. Retrieved July 1, 2018, from https://images.google.com/

[41] Walrus Scratching. Video. (31 July 2013). Retrieved July 30, 2018 from http://www.youtube.com/watch?v=zReIGuJR1J8

[42] G. Ren, X. Ni, M. Malik,. and Q. Ke. 2018. Conversational Query Understanding Using Sequence to Sequence Modeling. In *WWW '18. International World Wide Web Conferences*. 1715–1724.

Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering

Wenlong Sun*
wenlong.sun@louisville.edu
University of Louisville
Louisville, KY

Sami Khenissi*
sami.khenissi@louisville.edu
University of Louisville
Louisville, KY

Olfa Nasraoui*
olfa.nasraoui@louisville.edu
University of Louisville
Louisville, KY

Patrick Shafto*
patrick.shafto@rutgers.edu
Rutgers University - Newark
Newark, NJ

ABSTRACT

Recommender Systems (RSs) are widely used to help online users discover products, books, news, music, movies, courses, restaurants, etc. Because a traditional recommendation strategy always shows the most relevant items (thus with highest predicted rating), traditional RS's are expected to make popular items become even more popular and non-popular items become even less popular which in turn further divides the haves (popular) from the have-nots (unpopular). Therefore, a major problem with RSs is that they may introduce biases affecting the exposure of items, thus creating a popularity divide of items during the feedback loop that occurs with users, and this may lead the RS to make increasingly biased recommendations over time. In this paper, we view the RS environment as a chain of events that are the result of interactions between users and the RS. Based on that, we propose several debiasing algorithms during this chain of events, and evaluate how these algorithms impact the predictive behavior of the RS, as well as trends in the popularity distribution of items over time. We also propose a novel blind-spot-aware matrix factorization (MF) algorithm to debias the RS. Results show that propensity matrix factorization achieved a certain level of debiasing of the RS while active learning combined with the propensity MF achieved a higher debiasing effect on recommendations.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Collaborative search**.

KEYWORDS

Recommender Systems (RSs), Matrix Factorization, Active Learning (AL), Propensity

ACM Reference Format:
Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317303

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317303

1 INTRODUCTION

The goal of a RS is to, given the user's previous ratings, predict which items the user might like. Modern RSs generally aim to discover a constant relationship between users and items. This may lead to a situation in which users only see a narrow subset of the entire range of available recommendations, a phenomenon known as the 'filter bubble' [9]. The relationship between users and items is, however, a time dependent variable because the RS predicts that some items may not be of interest to the user, and therefore these items may never actually be seen by the user. This presents a significant problem for RSs: we might know why a user likes an item, but we do not know why an item is not-liked by the user. Is it not-liked by a user because the user does not like it, or is it simply because the user has not seen the item in the RS results? Furthermore, if we assume the RS will continue to recommend items to users based on biased ratings, and that users will respond to these recommendations, the RS will slowly learn to recommend increasingly similar items. In other words, the RS will begin to systematically limit the users' ability to discover more items [15]. In this paper, we propose to model how iterated biases evolve from the continuous user-RS feedback loop, develop a series of different debiasing strategies, and evaluate how these algorithms impact the predictive accuracy of the RS, as well as trends in the popularity distribution of items over time. We also propose a novel blind spot aware matrix factorization algorithm to debias the RS.

1.1 Objectives and Contributions

First, we argue that a RS is a chain of events in which users actively interact with the output of a RS. Next, we introduce several debiasing algorithms for RSs, particularly those based on MF, during this chain of events. To debias the RS interactions with users, we propose to use three algorithms:

1) a unified recommendation and active learning strategy (active recommendation) during the interaction between users and the RS algorithm, with the goal of reducing recommendation uncertainty, while at the same time ensuring the integrity of the algorithm's performance;

2) an exposure-based collaborative filtering recommendation model that is also combined with an active recommendation to further debias the RS;

3) a blind spot aware MF, which takes into account the blind spot inherent in the learning phase of the RS.

2 RELATED WORK

2.1 Debiasing Strategies

Work on debiasing RSs has been done from various perspectives. Hu et al. first proposed using an implicit feedback model to measure the level of confidence that a user will see an item [5]. This is similar to the modern notion of a propensity score [10], which is defined as the probability that an item will be seen by the user. Schnabel et al. [10] argued that recommending items in a RS is analogous to exposing a patient to a randomized treatment in a medical study, and proposed the introduction of a propensity-scored recommendation learning system. Liang et al. proposed that a propensity score matrix be calculated first, followed by a weighted MF based on the propensity score matrix [6, 7]. Abdollahpouri et al. proposed a fairness-aware regularization aiming to reduce popularity bias in recommender system [1]. Chanely et al. observed that algorithms increase homogeneity among users, thereby decreasing the utility gained by users of the system, and presented a simulation to illustrate how this effect occurs [2]. Yang et al. considered implicit feedback models instead of explicit feedback models in their study, and proposed ways to estimate a propensity matrix based exclusively on popularity, which essentially describes propensity as an estimator of the true probability distribution [17].

Singh et al. proposed a method to construct fair rankings among relevant items, positing that the problem originates at the ranking step (recommendation step) and not during the learning process [12]. Their methodology ranked the items based on a calculated utility, and added a fairness constraint based on the propensity score. Sinha et al. considered the RS as a feedback loop, assumed that user ratings are true prior to the feedback loop, and proposed a method to deconvolve this feedback mechanism, assuming that each feedback response follows a certain mathematic relationship to previous recommendations [13]. Nasraoui et al. considered the continuous interaction between learning algorithms and humans as a Markov Chain of event, and proposed several approaches to debias the learning algorithms [8]. Shafto and Nasraoui presented a cognitive foundation for the interaction between learning algorithms and humans, showing how recommendation algorithms may affect the behavior of humans within the *human-recommender* feedback loop, and proposed possible cognitive models to study and debias the interaction [11].

Unlike the above work, Our proposed strategies aim to reduce the iterated bias that occurs during the interactions between users and the RS without the introduction of strong assumptions. Moreover, in contrast to the aforementioned research, we propose algorithms that consider the RS to be a chain of events, and then focus on debiasing the iterated bias introduced by these interactions by using an estimated propensity score, with and without an active learning strategy. We also employ the Gini coefficient and the blind spot score to quantify how the interaction affects the users' ability to discover new items.

2.2 Propensity and Active Learning

We start by summarizing the notation used in this paper, which follows the notation in [7, 10]:

$R_{u,i}$: An integer which indicates the ratings of user u to item i.
$O_{u,i}$: A binary value which indicates that user u provided a rating for item i to the system, $[O_{u,i} = 1] \rightarrow [R_{u,i}\ is\ observed]$.
$P_{u,i}$: Propensity: The probability of observing an entry. $P_{u,i} = P(O_{u,i} = 1)$.
N_U: The total number of users.
N_i: The number of users who rated item i.

2.2.1 Propensity. Recommendations contain two sources of information: the items which the user can see, and the user's recorded preference toward those items. *Propensity* refers to the probability of observing $R_{u,i}$ [7, 10]. In a real-world application, what the user sees is highly subject to selection biases. For example, in a movie recommendation system, users watch and rate movies that they like, and rarely rate movies that they dislike since they may not have seen these movies. Another example would be an advertisement recommendation system that always shows ads that it believes to be of interest to the user. This bias is expected to deepen as a result of the iterative interaction between users and the recommendation outputs.

Recent research on RSs began to take into account the role of item propensity, where user exposure to an item in a RS is viewed as analogous to exposing a patient to a treatment in a medical study [7, 10]. As a result, propensity indicates how probable it is that a new treatment can or will be exposed to a patient. In both cases, the studies try to infer the effect based on current results (whether it be the effect of a new treatment, or a new item in a RS).

Existing approaches for RSs generally under-weight items that are not rated in the system. It is, however, difficult to determine whether an item is not rated by a user because the user does not like the item, or because the user has not seen the item as a result of the inherent selection bias in the RS [7]. The main idea behind a propensity-based MF is to under-weight the unrated item for recommendation by bringing into the objective function of the model an Inverse Propensity Score, as follows [10]:

$$\underset{V,M}{argmin} \sum_{O_{u,i}=1} \frac{||R - V^T M||^2}{P_{u,i}} + \lambda(||V||_F^2 + ||M||_F^2). \quad (1)$$

Here $P_{u,i}$ represents the probability that a user u will see item i, and is also referred to as propensity score. V and M are the two latent factors in the MF.

Estimating propensity. A simple way of estimating propensity is to use a popularity score [7]. This assumes that $O_{u,i}$ follows a Bernoulli distribution, i.e., $O_{u,i} \sim Bernoulli(\rho_i)$. Note that the propensity score is fixed across users in this case, i.e., $P_{u,i} = \hat{\rho}_i$. Given a rating matrix, the popularity of an item is the proportion of items exposed to certain users among all the users (essentially, the proportion of users who have rated the item relative to all users).

Another way of estimating propensity is to assume that $O_{u,i}$ follows a Poisson distribution [7], or $O_{u,i} \sim Pois(\pi_u^T \gamma_i)$. π_u and γ_i are the two latent factors of the Gamma prior. Given a ratings matrix, this method assigns a value of 1 to an item that was rated

by a user in the observational matrix O, and 0 if it is not rated. By factoring this observational matrix, we get the propensity scores of all user and item pairs (see Eq. 2)

$$P_{u,i} = 1 - P(O_{u,i} = 0|\pi_u, \gamma_i) \approx 1 - exp\left\{-E[\pi_u^T \gamma_i]\right\}. \quad (2)$$

2.2.2 Active Learning . Active Learning (AL) is a special case of semi-supervised learning in which the system has the ability to interactively prompt users to label (or rate) items in order to improve the accuracy of the model [3]. One of the advantages of AL is that the targeted knowledge the system acquires helps accelerates the speed in which the system learns the model. One way to implement AL is to actively prompt users to rate items that have been underweighted by the RS in order to improve the quality of the ratings of this subset of items. In a rating system with range from R_{min} to R_{max} , the active learning can be formalized directly as follows [18]: Select the next item x_{act} that satisfies

$$x_{act} = \underset{x_i}{argmin}[\theta - \hat{y}|x_i]. \quad (3)$$

Here, $\hat{y}$ is the rating predicted by the RS given an item x_i. θ controls the degree of active learning, ranging from the midrange rating of $0.5(R_{min} + R_{max})$ to the maximum rating R_{max}. It can be seen that $\theta = R_{max}$ recovers pure recommendation (select the most relevant item, hence the item with highest predicted rating), while $\theta = 0.5(R_{min} + R_{max})$ recovers pure active learning (select the item with most uncertain relevance to the user based on the predicted rating, hence an item that is far from both being very relevant ($\hat{y} = R_{max}$) and very non-relevant ($\hat{y} = R_{min}$)). Cognitive experiments have shown that an active recommendation system can cover a wider choice of items, while maintaining the accuracy of the results [18].

3 PROPOSED METHODS TO DEBIAS RECOMMENDER SYSTEMS

We first introduce our interactive recommender system framework, which considers a RS as a continuous chain of events. First the initial RS suggests items to each user based on the initial training ratings and it is assumed that the users have 100% agreement with the recommendation. We then retrieve the true ratings from our masked ratings and add to the new training ratings. After that, a new recommendation based on new training data will be issued. This interaction will continue until a maximal number of iterations is reached. Algorithm 1 shows the details of our interactive recommender system with human in the loop.

We propose several recommendation strategies based on well-known algorithms to simulate real-life user-recommendation system interaction.

Conventional MF:
This model is trained using conventional MF (same as Eq. 1 with $P_{u,i} = 1$), and the system always selects the top predicted item for each user, and adds it to the next (new) training set. In other words, there is no active learning.

Conventional MF + Active Learning:
This model is trained using conventional MF (Eq. 1 with $P_{u,i} = 1$), but the system selects the active recommendation items with $\theta = 4.5$ for each user in Eq. (3). $\theta = 4.5$ is chosen to be between $\theta = 0.5(R_{min} + R_{max}) = 3$ (pure AL) and $\theta = R_{max} = 5$ (pure recommendation).

Popularity Propensity MF:
This model is trained with propensity MF (Eq. 1) [10]. Here the propensity $P_{u,i}$ is estimated based on popularity. The system always selects the top predicted item for each user, and adds it to the next (new) training set. In other words, there is no active learning.

Popularity Propensity MF + Active Learning:
This model is trained with propensity MF (Eq. 1) [10]. Here the propensity $P_{u,i}$ is estimated using popularity. The system selects the active recommendation items with $\theta = 4.5$ for each user in Eq. (3).

Poisson Propensity MF:
The model is trained with propensity MF (Eq. 1). Here the propensity $P_{u,i}$ is estimated based on Poisson MF (2) [7] on the exposure matrix. The system always selects the top predicted item for each user, and adds it to the next (new) training set. In other words, there is no active learning.

Poisson Propensity MF + Active Learning:
The model is trained with propensity MF (Eq. 1). Here the propensity $P_{u,i}$ is estimated using Poisson MF (Eq. 2) [7] on the exposure matrix and the system selects the active recommendation items using (Eq. 3) with $\theta = 4.5$ for each user (again chosen to be between $\theta = 0.5(R_{min} + R_{max}) = 3$ (pure AL) and $\theta = R_{max} = 5$ (pure recommendation)).

Algorithm 1: Interactive Recommendation System with the Human-Recommender System Feedback Loop Debiasing Mechanism

Data: Rating matrix R'_{ui}, λ, Learning rate η, $MAX_{interation}$, Iterations=0, Size of selection
Result: MAE, RMSE, Gini Coefficient
The system trains the initial Matrix Factorization model and computes predictions $\hat{R}$;
while *Iterations < Max Feedback Loop Iterations* **do**
  1 for all users u in the system {
    1.1. The system selects top-N items to recommend from the predicted ratings $\hat{R}$ based on a specialized recommending strategy;
    1.2. User u picks the selected top-N items and gives rating $R^{new}_{u,i}$ (from ground-truth complete data) for each item i;
  }
  2. The system records the popularity $P_i = N_i/N_U$ after the new ratings are taken in.
  3. The system records the metrics such as the RMSE and the Gini index of the popularity given the current rating matrix;
  4. The system retrains the model with the new rating matrix using *steps* gradient descent updates (Eq. 4 or Eq. 8 depending on the recommendation strategy chosen, and recomputes the predictions).
  5. Iterations++
end

In this preliminary work, we use both popularity and Poisson matrix factorization (PMF) to estimate the propensity. To minimize the objective function, we use stochastic gradient descent, which has been used successfully to solve MF with big datasets. For a given training rating r_{ij}, the updates for V_u and M_i can be shown to be:

$$\begin{aligned} V_u &\leftarrow V_u + \eta(2e_{ui}M_i - \lambda V_u) \\ M_i &\leftarrow M_i + \eta(2e_{ui}V_u - \lambda M_i). \end{aligned} \tag{4}$$

Here $e_{ui} = \frac{(\hat{r}_{ui} - V_u^T M_i)}{P_{u,i}}$, $\hat{r}_{ui}$ is the predicted rating and η is the learning rate for gradient descent. With a proper choice of step size, gradient descent converges to a local minimum. Propensity $P_{u,i}$ is computed as follows:

$$P_{u,i} = \begin{cases} 1.0 & for\ Conventional\ MF \\ N_i/N_U & for\ Popularity\ Propensity\ MF \\ 1 - P(O_{u,i} = 0 | \pi_u, \gamma_i) & for\ Poisson\ Propensity\ MF \end{cases} \tag{5}$$

This means that $P_{u,i} = 1$ for the first two (Conventional MF, with or without AL) recommendation strategies. $P_{u,i}$ is estimated using popularity for Popularity Propensity MF (with or without AL) strategies and it is estimated using Eq. 2 for Poisson Propensity MF (with or without AL) strategies.

In addition to the above proposed debiasing strategies above, a seventh algorithm called **Blind Spot Aware Matrix Factorization** is introduced. In this paper, we define the blind spot size as the number of item with a predicted ratings $\hat{R_{u,i}}$ that is smaller than a threshold δ , i.e., $\mathbf{D}_{\delta}^{\mathbf{u}} = \{\mathbf{i} \in I \mid \hat{R}_{u,i} < \delta\}$. Note that because each user has their own blind spot, we define a threshold for each user. The threshold is used to set a percentile cut-off for each user, 95% in our experiments. Therefore, We define the blind spot for user u as

$$\mathbf{D}_{\epsilon}^{\mathbf{u}} = \{\mathbf{i} \in I \mid \hat{R}_{u,i} < max_{u,i}(\hat{R}_{u,i}) * \epsilon\}. \tag{6}$$

Here ϵ is a cut-off which controls the threshold.

Our proposed Blind Spot Aware MF tries to limit the blind spot when trying to optimize the cost function of the conventional matrix factorization. The cost function for blind spot aware MF is as follows:

$$\begin{aligned} J = \sum_{u,i \in R} ||r_{u,i} - V_u^T M_i||^2 + \frac{\lambda}{2}(||V_u||^2 + ||M_i||^2) \\ + \underbrace{\frac{\beta}{2}||V_u - M_i||^2}_{Blind\ \ Spot\ \ Aware\ \ Term}. \end{aligned} \tag{7}$$

To minimize the objective function, we use stochastic gradient descent, which has been used successfully to solve MF for CF with big datasets. For a given training rating r_{ij}, the updates for V_u and M_i can be shown to be:

$$\begin{aligned} V_u &\leftarrow V_u + \eta(2e_{ui}M_i - \lambda V_u - \beta(V_u - M_i)) \\ M_i &\leftarrow M_i + \eta(2e_{ui}V_u - \lambda M_i + \beta(V_u - M_i)). \end{aligned} \tag{8}$$

Here $e_{ui} = \hat{r}_{ui} - V_u^T M_i$, $\hat{r}_{ui}$ is the predicted rating, and η is the learning rate for gradient descent. With a proper choice of step size, gradient descent converges to a local minimum.

Intuition Behind Blind Spot Aware MF: Conventional Matrix Factorization aims to predict ratings by approximating the existing ratings by the dot product between user and item. The RS tries to differentiate different items for each user as much as possible (see Figure 1). Given the same proportional range of similarity, blind spot aware MF aims to recommend more items to explore (with high relevance scores). Conventional MF has more limited choices to recommend to users given the predicted ratings range $[0.8\hat{R}_{max}, \hat{R}_{max}]$, where $\hat{R}_{max}$ is the maximum predicted rating for a user.

On the other hand, Blind Spot Aware MF tries to match the existing rating, while bringing items close to each other so that a user has a higher chance to explore more items, so that the overall blind spot size is decreased (see figure 1).

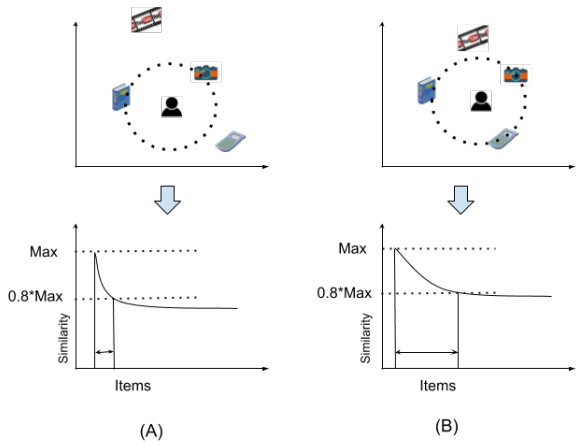


Figure 1: Conventional Matrix Factorization vs. Blind Spot Aware Matrix Factorization. Figure (A) indicates Conventional Matrix Factorization, in which the algorithm aims to find items that are close to users through differentiation. The top of Figure (A) indicates how items are distributed around a user in the latent space under Conventional MF, the bottom of (A) shows the similarity between the user and all items in latent space. Figure (B) shows how the proposed Bias-aware Matrix Factorization finds items close to users while keeping the items that are close to each other in the latent space. The top of Figure (B) indicates how items are distributed around a user in the latent space , the bottom of (B) shows the similarity between the user and all items in latent space.

4 EXPERIMENTS

4.1 Data Set

We use item response theory to generate a sparse rating matrix $\boldsymbol{R}_{u,i}$ using the model proposed in [4]. Assume a_u to be the center of user u's rating scale, and b_u to be the rating sensitivity of user u. Finally let t_i be the intrinsic score of item i. We generate a user-item rating matrix as follows,

$$\boldsymbol{R}_{u,i} = L[a_u + b_u t_i + \eta_{u,i}], \tag{9}$$

where $L[\omega]$ is the discrete level function, assigning a score in the range $R_{min} = 1$ to $R_{max} = 5 : L[\omega] = \max(\min(\text{round}(\omega),$

5), 1) and $\eta_{u,i}$ is a noise parameter. In our experiments, we draw $a_u \sim N(3.4, 1)$, $b_u \sim N(0.5, 0.5)$, $t_u \sim N(0.1, 1)$, and $\eta_{u,i} \sim \epsilon N(0, 1)$; where N is a standard normal density, and ϵ is a noise parameter, we set up $\epsilon = 0.5$. We generate a rating matrix R with 500 users and 500 items, therefore we have 250,000 ratings in total.

4.2 Evaluation

In order to assess the accuracy of the prediction of the RS during the interactive recommendation, we compute the Root Mean Square Error (RMSE). To check the impact of the debiasing mechanism, we compute the Gini coefficient of the popularity scores of all items. The Gini coefficient is used to measure the inequality of a distribution [14]. Let P_i be the popularity of each item after training the model. For a population with n values P_i, $i = 1$ to n, that are indexed in non-decreasing order ($P_{(i)} \leq P_{(i+1)}$), the Gini coefficient can be calculated as follows [14]:

$$G = \left(\frac{\sum_{i=1}^{n}(2i - n - 1)P_{(i)}}{n\sum_{i=1}^{n} P_{(i)}}\right). \tag{10}$$

The higher the Gini coefficient, the more unequal are the values in the dataset. The Gini coefficient of the popularity of items shows how the recommendation system's output affects the exposure distribution of items. Traditional RS's are expected to make popular items become even more popular and non-popular items become even less popular because a traditional recommendation strategy always shows the most relevant items (thus with highest predicted rating), which further divides the haves (popular) from the have-nots (unpopular).

4.3 Method

We first randomly selected 25 items for each of the 500 users from the completed rating matrix and started our initial training. We then select 20% of the ratings as the test set. Note that the testing set is later fixed, and will not be changed. All other ratings are considered as candidate ratings: they are used to simulate the feedback loop of RS and human interaction, and will be added based on the selection mechanism of the recommendation strategy in each iteration/loop. Figure 2 shows the approach for splitting our data. We record the RMSE and the Gini coefficient of predicted item popularity of each testing set.

In each simulation of a feedback loop, we use one of the recommendation strategies listed in Section 3 to recommend items to each user. After that, we simulate the users' response by assuming that they responded to items that are recommended and provide the true ratings for the top-N ratings (top-N=10). After new ratings are taken in, we update the recommender system. We then simulate runs of *Max Feedback Loops* = 20 iterations in Algorithm 1 where a single iteration (or loop) consists of the algorithm providing a recommendation, the user labeling the recommendation, and the algorithm updating its model of the user's preferences.

For each of the recommendation strategies in section 3, we set the number of items (top-N) selected after each recommendation to 10. For the blind spot aware MF and conventional MF with and without AL models, we use stochastic gradient descent to optimize the objective function. For all the Matrix Factorization methods, we set the dimension of the latent space to 20. For all the four propensity based MF algorithms, we optimize the objective function following coordinate gradient descent update rules to learn the model [16]. For all gradient descent optimization updates, we set the number of learning iterations *steps* = 200, regularization weight $\lambda = 0.02$ and learning rate $\eta = 0.002$ for training the matrix factorization model. All the results are from the average of 10 independent runs.

We also report the results for random selection as baseline. This means that the items are selected randomly by the user to rate after the recommendations are made by using the Conventional MF model (meaning essentially an *open loop*).

For the Blind Spot Aware MF 7, We determine the blind spot aware term weight coefficient β by grid search, i.e., running the matrix factorization (Eq. 7) with different $\beta = [0.002, 0.02, 0.05, 0.1, 0.5, 0.8]$ on the initial rating data (before the completion). Finally, we choose β as 0.2 for the Blind Spot Aware term.

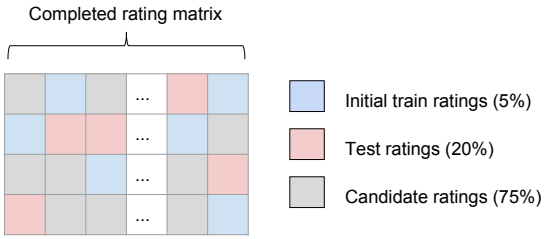


Figure 2: The splitting of the completed rating matrix. The completed rating matrix is split into 3 parts: 1) Initial ratings; 2) Test ratings; and 3) Candidate ratings. The candidate ratings will be added to the training ratings when queried by the system.

4.4 Results

Figure 3 shows the RMSE of each recommendation strategy during the iterative recommender system for the training ratings. The conventional MF and the conventional MF with AL show similar trends in the RMSE (identical on the plot). However, the conventional MF with random selection starts with a high RMSE, but this decreases with each iteration when more ratings are collected. All propensity based MF methods, with or without AL, show a similar trend for the training RMSE as well as the blind spot aware MF. At iteration 1, all six algorithms have the same initial training ratings. The reason why they have large differences on training RMSE lies in the fact that all the propensity MF strategies do not minimize the RMSE; instead they minimize the primary loss term consisting of the square error loss between the prediction $\hat{r}_{ui}$ and the true rating r_{ui}, inversely weighted by $\frac{1}{P_{u,i}}$ (see Eq. 1). On the other hand, conventional MF directly minimizes RMSE, i.e. does not weight this error in the primary loss term (which is equivalent to setting $P_{u,i} = 1$). Note that random baseline means that the items are selected randomly after the recommendations (meaning essentially an open loop).

We also record the RMSE on the testing ratings, as shown in Figure 4. The RMSE for the testing set increases dramatically in the early stages for all propensity based MFs, but then drops to a low level. On the other hand, random selection and conventional MF approaches have a strictly decreasing trend in RMSE with more

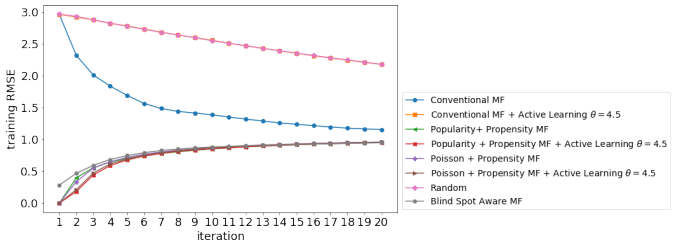


Figure 3: Training primary loss term with different debiasing mechanisms for each iteration on synthetic data. The x-axis represents the feedback loop iteration number, and the y-axis represents the RMSE for the training set. The conventional MF and the conventional MF with AL show a similar trend during each iteration. All four propensity based MFs also show a similar trend as well as the blind spot aware MF. Note that random baseline means that the items are selected randomly after the recommendations (meaning essentially an open loop).

iterations, however with a higher value compared to other algorithms. The blind spot aware MF algorithm has the lowest testing error compared to other algorithms.

Gini Coefficient. As stated in section 2, we computed the Gini coefficient of the item popularity after each feedback loop iteration to assess how different debiasing mechanisms affect the distribution of popularity. A higher Gini index indicates more heterogeneity of the popularity, essentially meaning a bigger popularity divide between the items, leading to a wider gap between the haves and the have-nots. Figure 5 shows the Gini coefficient distribution. Conventional MF increases the inequality of the popularity of items, which indicates that the conventional MF will boost some popular items. On the other hand, the pure Propensity based MF has a high Gini coefficient at an early stage which then decreases with each feedback loop iteration. Propensity based MF with AL results in a low Gini coefficient across all feedback loop iterations. Random selection appears to have the lowest Gini coefficient since all items

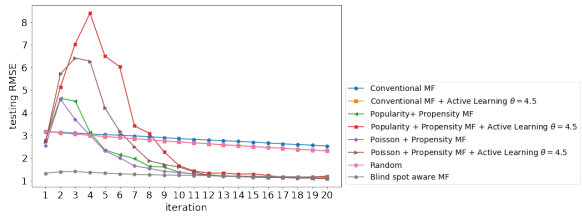


Figure 4: Testing RMSE with different debiasing mechanisms for each iteration on synthetic data. The x-axis represents the feedback loop iteration number, and the y-axis represents the RMSE for the testing set. Conventional MF, with and without AL, show a similar trend during each iteration, and they are similar to random selection. All four propensity based MFs also show a similar trend with lower RMSE, as well as the blind spot aware MF.

have the same probability of being explored. Blind Spot Aware MF has a lower Gini coefficient compared with Conventional MF.

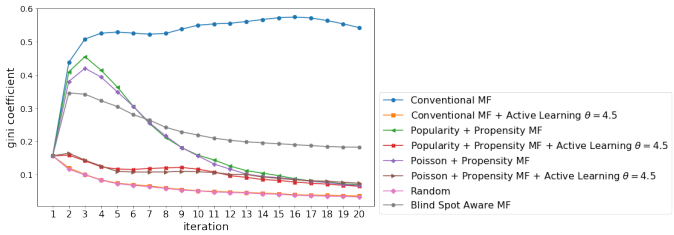


Figure 5: Gini coefficient vs. feedback loop iteration on the synthetic data. The Gini coefficient increases for the Popularity Propensity MF without AL, but then quickly decreases. On the other hand, the Gini coefficient score of the Propensity MF with AL continues to decrease. Note that a higher Gini index indicates more heterogeneity of the popularity, essentially meaning a bigger popularity divide between the items.

Based on the observed trends of the RMSE and the Gini coefficient analysis, we conclude that: 1) Propensity MF achieves a significant level of debiasing on the recommendations in terms of Gini coefficient ($p_{value} < 0.05$); 2) Active Learning combined with Propensity MF results in low testing RMSE and a high debiasing effect on recommendations; 3) Conventional MF increases the popularity bias which eventually affects the ability to discover new items, but decreases the bias when combined with an Active Learning strategy; 4) Blind Spot Aware MF achieves one of the lowest testing RMSE and decreases popularity bias compared to Conventional MF.

5 CONCLUSIONS

Recommender systems introduce bias during the interactive feedback loop with users over time. In this paper, we introduce an interactive framework to simulate the feedback loop that is created by the chain of events generated when a recommender system interacts with users over time. Based on this framework, we proposed several debiasing algorithms based on existing and new techniques to use during this chain of events. We also proposed a blind spot aware matrix factorization algorithm which takes into account the blind spot score when trying to learn the recommendation model.

Note that in this paper, we do not focus on improving collaborative filtering algorithms (Matrix Factorization) for recommender systems by studying user feedback. Instead, our goal is to simulate the interaction between users and the recommender system and to debias the recommender system during the interaction. Our results showed that propensity based MF achieved a certain level of debiasing of the RS while active learning combined with the propensity MF achieved a higher debiasing effect on recommendations. Our proposed blind spot aware matrix factorization also achieved a certain level of debiasing of the RS. One limitation of this study is that we assume that users totally agree with the recommendations in each iteration, and provide feedback. In real-life, users might not agree with recommendations. Thus future study could use a more realistic user reaction model.

ACKNOWLEDGEMENT

This work was supported by National Science Foundation grant NSF-1549981.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 42–46.

[2] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2017. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *arXiv preprint arXiv:1710.11214* (2017).

[3] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research* 4, 1 (1996), 129–145.

[4] David F Gleich and Lek-heng Lim. 2011. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 60–68.

[5] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 263–272.

[6] Dawen Liang, Laurent Charlin, and David M Blei. [n. d.]. Causal Inference for Recommendation. ([n. d.]).

[7] Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. 2016. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 951–961.

[8] Olfa Nasraoui and Patrick Shafto. 2016. Human-algorithm interaction biases in the big data cycle: A markov chain iterated learning framework. *arXiv preprint arXiv:1608.07895* (2016).

[9] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. ACM, 677–686.

[10] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).

[11] Patrick Shafto and Olfa Nasraoui. 2016. Human-recommender systems: From benchmark data to benchmark cognitive models. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 127–130.

[12] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. *arXiv preprint arXiv:1802.07281* (2018).

[13] Ayan Sinha, David F Gleich, and Karthik Ramani. 2016. Deconvolving feedback loops in recommender systems. In *Advances in Neural Information Processing Systems*. 3243–3251.

[14] Alan Stuart, J Keith Ord, and Sir Maurice Kendall. 1994. *Distribution theory*. Edward Arnold; New York.

[15] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. 2018. Iterated Algorithmic Bias in the Interactive Machine Learning Process of Information Filtering. (2018).

[16] Stephen J Wright. 2015. Coordinate descent algorithms. *Mathematical Programming* 151, 1 (2015), 3–34.

[17] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased Offline Recommender Evaluation for Missing-Not-At-Random Implicit Feedback. (2018).

[18] Scott Cheng-Hsin Yang, Jake Alden Whritner, Olfa Nasraoui, and Patrick Shafto. [n. d.]. Unifying recommendation and active learning for human-algorithm interactions. ([n. d.]).

Workshop on Hypermedia Multi-Agent Systems (HyperAgents) Chairs' Welcome

It is our great pleasure to welcome you to the *First Workshop on Hypermedia Multi-Agent Systems (HyperAgents 2019)* at *WWW 2019*. HyperAgents 2019 invites researchers and practitioners to share their thoughts and vision on what the future of a hypermedia-driven Web for both people and autonomous agents will be.

Hypermedia is increasingly used in Web service design, particularly in *Web of Things* and *Linked Data* systems: the dynamic and open nature of these systems requires components to be deployed and to evolve independently from one another, which makes the use of static Web APIs that are manually integrated by developers simply impractical (in any such systems). This evolution raises new challenges: to *discover*, *consume*, and *integrate* hypermedia APIs at runtime, clients have to become increasingly autonomous in pursuit of their design goals. Such autonomous systems have been studied to a large extent in research on *multi-agent systems (MAS)*. To consolidate the evolution of hypermedia APIs, it is now necessary to have comprehensive discussions on integrating hypermedia systems and MAS, which we call *Hypermedia MAS*.

For this first edition, we received 4 submissions, out of which the program committee has selected 3 papers for presentation at the workshop. In addition, the workshop agenda will feature two keynote presentations from prominent speakers, Mike Amundsen[1] and Florian Michahelles (Siemens), an invited contribution from Alessandro Ricci (University of Bologna), and breakout sessions meant to encourage open discussions and ideation at the workshop.

We would like to thank the authors for their submissions and to the program committee for providing substantial feedback to authors.

We look forward to welcome you to an exciting event!

Andrei Ciortea
University of St. Gallen, Switzerland

Simon Mayer
University of St. Gallen, Switzerland

Fabien Gandon
Inria, Université Côte d'Azur, CNRS, I3S, France

Olivier Boissier
Mines Saint-Étienne, Université de Lyon, LabHC, CNRS, France

Workshop on Hypermedia Multi-Agent Systems (HyperAgents) Organization

[1] An internationally known author and speaker, Mike has authored the book "RESTful Web Clients: Enabling Reuse Through Hypermedia" (O'Reilly Media, 2017).

Workshop Chairs: Andrei Ciortea *(University of St. Gallen, Switzerland)*
Simon Mayer *(University of St. Gallen, Switzerland)*
Fabien Gandon *(Inria, Université Côte d'Azur, CNRS, I3S, France)*
Olivier Boissier *(Mines Saint-Étienne, Université de Lyon, LabHC, CNRS, France)*

Program Committee: Mihai Bâce *(ETH Zürich, Switzerland)*
Abraham Bernstein *(University of Zurich, Switzerland)*
Mike Blackstock *(University of British Columbia, Canada)*
Olivier Boissier *(MINES Saint-Étienne, France)*
Rafael Bordini *(Pontifícia Universidade Católica do Rio Grande do Sul, Brazil)*
Jean-Paul Calbimonte *(HES-SO, Switzerland)*
Amit Chopra *(Lancaster University, UK)*
Andrei Ciortea *(University of St. Gallen, Switzerland)*
Rem Collier *(University College Dublin, Ireland)*
Olivier Corby *(Inria, France)*
Amal El Fallah-Seghrouchni *(Sorbonne Université, France)*
Fabien Gandon *(Inria, Université Côte d'Azur, CNRS, I3S France)*
Kimberly Garcia *(Siemens, USA)*
Christophe Guéret *(Accenture, Ireland)*
Dominique Guinard *(EVRYTHNG, UK)*
Amélie Gyrard *(Wright State University, USA)*
Armin Haller *(Australian National University, Australia)*
Siegfried Handschuh *(University of St. Gallen, Switzerland)*
Andreas Harth *(University Erlangen-Nuremberg and Fraunhofer, Germany)*
Dominique Hazël-Massieux *(W3C, France)*
Jomi Hübner *(Universidade Federal de Santa Catarina, Brazil)*
Matthias Klusch *(DFKI, Germany)*
Michael Koster *(SmartThings, USA)*
Matthias Kovatsch *(Huawei, Germany)*
Maxime Lefrançois *(MINES Saint-Étienne, France)*
Viviana Mascardi *(University of Genova, Italy)*
Friedemann Mattern *(ETH Zürich, Switzerland)*
Simon Mayer *(University of St. Gallen, Switzerland)*
Florian Michahelles *(Siemens, USA)*
Franck Michel *(Université Côte d'Azur, France)*
Petros Papapanagiotou *(University of Edinburgh, UK)*
Cesare Pautasso *(University of Lugano, Switzerland)*
Alessandro Ricci *(University of Bologna, Italy)*
Till Riedel *(KIT, Germany)*
Michael Rovatsos *(University of Edinburgh, UK)*
Josh Siegel *(Michigan State University, USA)*
Munindar Singh *(North Carolina State University, USA)*
Alexandru Sorici *(University "Politehnica" of Bucharest, Romania)*
Ruben Verborgh *(Ghent University, Belgium)*

Erik Wilde (Good *API, Switzerland)*
Antoine Zimmermann *(MINES Saint-Étienne, France)*

Reviewers: Arthur Casals *(Universidade de São Paulo, Brazil)*

THE WEB CONFERENCE

MAMS: Multi-Agent MicroServices*

Rem W. Collier
University College Dublin
Dublin, Ireland
rem.collier@ucd.ie

Eoin O'Neill
University College Dublin
Dublin, Ireland
eoin.o-neill.3@ucdconnect.ie

David Lillis
University College Dublin
Dublin, Ireland
david.lillis@ucd.ie

Gregory M.P. O'Hare
University College Dublin
Dublin, Ireland
gregory.ohare@ucd.ie

ABSTRACT

This paper explores the intersection between microservices and Multi-Agent Systems (MAS), introducing the notion of a new approach to building MAS known as Multi-Agent MicroServices (MAMS). Our approach is illustrated through a worked example of a Vickrey Auction implemented as a microservice.

CCS CONCEPTS

• **Information systems → RESTful web services**; • **Computing methodologies → Multi-agent systems**; • **Software and its engineering →** *Organizing principles for web applications.*

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
Rem W. Collier, Eoin O'Neill, David Lillis, and Gregory M.P. O'Hare. 2019. MAMS: Multi-Agent MicroServices. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316509

1 INTRODUCTION

As a research field, Multi-Agent Systems (MAS), has existed since the late 1980s. Since its inception, notions such as loose-coupling, distribution, reactivity and isolated (local) state have been core concepts [14, 45]. While the notion of an agent has been widely researched and a large range of tools and programming languages developed [1, 3, 38, 40], there has been little real adoption of these languages and tools within industry [8].

At the same time, industry has evolved through a range of enterprise paradigms and models, that have slowly shifted from monolithic centralised systems towards highly-decentralised systems that increasingly exhibit properties that were once seen as differentiators for MAS. Currently, it is the era of the microservice [20]; a model in which systems are built from small, loosely-coupled services that maintain their own independent state [13]. Collectively, these services are deployed within an ecosystem of tools and components that facilitate rapid and agile development techniques; are easy to extend; and support automated management of fault tolerance and scaling [46].

This paper argues that microservices represent a potential point of convergence between modern software engineering and MAS. It adopts the position that an agent can be viewed as a type of microservice that can be deployed seamlessly within any microservice ecosystem. From an external perspective, such a microservice would be indistinguishable from other standard microservices; with all interaction being directed through a uniform interface. This would allow the developers of such a service to leverage existing tools and components to deliver a MAS that is founded upon, and closely aligned with, industry accepted tools and platforms.

2 FROM MICROSERVICES TO MULTI-AGENT SYSTEMS

Microservices represent the state of the art for large-scale software development. While not a one-size-fits-all solution, they have clearly demonstrated their potential as a tool for building systems at scale. We believe that microservices have many similarities with MAS. In this section, we examine both approaches by exploring how the principles of microservices relate to MAS and vice-versa.

Microservices [20] are a realisation of the established Service-Oriented Architecture (SOA) design style [46]. Due to its popularity, many definitions of microservices pervade the web. In an effort to be complete, we have picked two. Firstly, we consider the definition of [13] who argues that the core principles of microservices are: Bounded Context, Size, and Independence. Secondly, we consider [46] who argues that one of the basic tenets of microservices is their adherence to the IDEAL (Isolated state, Distribution, Elasticity, Automated management and Loose coupling) design principles [15], a set of guiding principles for the design of all cloud-based software.

Table 1, presents a combined summary of these two definitions that reflects our attempt to understand how the principles of microservices relate to MAS. Due to their similarity, Independence and Loose Coupling have been combined.

Overall, the comparison demonstrates that there are many commonalities between these two approaches: isolated state, distribution, elasticity and loose coupling seem to be equivalent. From an agent perspective, isolated state and loose coupling arise from the view of agents as autonomous decision-makers [45] and the

*Produces the permission block, and copyright information

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316509

Table 1: Comparing Microservice principles to MAS

Principle	Microservices	MAS
Bounded Context	A microservice represents a single piece of business functionality.	An agent can play a single or multiple roles in a system.
Size	Microservices should be small enough to ensure maintainability and extensibility.	Size/complexity is not an issue in MAS research and often depends on the target domain.
Isolated State	Sharing of state information is minimised across services.	State is local and private to an agent. This is often viewed as essential for an agents autonomy [ref].
Distribution	Services are spread across multiple nodes.	Agents are logically distributed, but it is also expected that they will be spread over multiple nodes.
Elasticity	The application is designed to allow addition and removal of required resources at runtime.	The ability to add/remove agents at runtime is a central feature of MAS.
Automated Management	Management operations like failure handling and scaling are automated.	Management operations are not central to agents, but are sometimes considered.
Loose Coupling	Systems are decomposed into loosely coupled sets of highly-cohesive colocated services.	Agents are autonomous and loosely-coupled problem solvers.

distribution perspective comes from the view of agents as social entities [45] that can be considered "...a loosely-coupled network of problem solvers..." [14]. Elasticity, in terms of the ability to create or destroy agents at run-time has been a central concept in MAS research since its inception, for example [32].

The bounded context and size principles seem less consistent. In our opinion, this is a consequence of how each approach has emerged. Microservices are a practical response to the limitations of monolithic applications and the complexity of other service models. In contrast, MAS is dominated by research and encompasses a diverse range of approaches. This has resulted in less emphasis being placed on such issues. Automated management is an area that has seen some attention in MAS research, often under the guise of Autonomic Computing [24]. However, such techniques are rarely considered key in MAS research.

Another key tenet listed by [46] is polyglot programming - the argument that developers can and should use appropriate programming paradigms (e.g. Java, Python, Matlab) for each microservice.

Table 2: Comparing MAS principles to Microservices

Principle	Microservices	MAS
Autonomy	Microservices operate without the direct intervention of humans and have some kind of control over their internal state (and actions?).	Agents operate without the direct intervention of humans and have some kind of control over their actions and internal state.
Social Ability	Interaction between microservices is typically achieved using messages based on RESTful APIs and HTTP.	Agents interact with other agents using some kind of Agent Communication Language.
Reactivity	Microservices respond to incoming HTTP requests in a timely fashion.	Agents perceive their environment and respond in a timely fashion to changes that occur in it.
Proactivity	Microservices do not take the initiative.	Agents don't just respond - they take the initiative.

To offer an opposing analysis, we consider how well microservices conform to agnet concepts. For this, we choose the classic weak notion of agency espoused by Wooldridge and Jennings [45] which argues that an agent is any entity that demonstrates autonomy, social ability, reactivity and proactivity. Table 2 compares how these features apply to both microservices and agents, with the MAS definitions taken directly from their paper.

Microservices seem to meet the autonomy, social ability and reactivity principles quite well. The differences revolve mainly around the fact that agents are situated entities, while microservices are not. From an MAS perspective, reactivity is closely related to the agents' environment. From a Microservices perspective, reactivity is simply the ability to respond effectively to incoming HTTP Requests. While they are clearly not the same level of complexity, if we view a microservices environment to be the set of messages it receives, then clearly it is reactive.

The view of microservices as entities that react just to messages is one that is being challenged in the context of Web of Things style applications [7] where things are represented as microservices that expose state and enable remote action [23]. While this reduces the conceptual gap for autonomy, it is not reflected in reactivity. This is because microservices principles are concerned primarily with defining external rather than internal attributes. As such, microservices do not consider how their internal state is realised and maintained, just that the state exists, is represented as a set of resources, and can be accessed/modified though a uniform (REST) interface. It is this same consideration that affects the proactivity principle - microservices principles do not cover proactivity because it is an internal quality. Finally, while social ability might seem compatible, it is less so when reflecting on the core objective of each approach: microservices aim to directly expose

(data) resources in a fine grained way, while agents seek to expose knowledge (state) and capability (action).

3 MULTI-AGENT MICROSERVICES (MAMS)

This section introduces *Multi-Agent MicroServices (MAMS)*, a class of system that is comprised of Agent-Oriented MicroServices (AOMS), and Plain-Old MicroServices (POMS). AOMS are microservices that are built using MAS technologies which are exposed through a well-defined interface modelled as a set of REpresentational State Transfer (REST) [16] [17] endpoints Such a system is an essential step in achieving our longer-term goal of realising Hypermedia MAS [7] - agent-based systems that are able to discover, consume and integrate hypermedia services which act as an enabler for Semantic Web and Linked Data systems [36, 41]

3.1 Agents and Microservices

In this section, we consider the relationship between agents and microservices. To do this, we first reflect on the relationship between resources and microservices that is promoted through REST.

Conceptually, REST promotes a view in which all resources are isolated as distinct services. In practice, resources are often co-located to minimize the impact of issues such as `chattiness` [39], which occurs where resources have been overly decomposed resulting in an abundance of network calls. This relates to the issue of Bounded Context discussed in section 2. The result of this is that a single microservice may often play host to multiple resources.

How does this relate to agents? Conceptually, agents are not just resources; they are complex decision-making entities that reason about their actions. On the other hand, agents rarely work in isolation, it is a common model that groups of agent work closely together through some form of collaborative process that is often realised through message-passing (e.g. auctions). As such, it is clear that such agents may suffer from similar issues, such as, chattiness. As with resources, it may make sense to co-locate such agents within a single microservice.

What about the relationship between agents and resources? Although there is a clear difference between agents and resources, we adopt a view that an agent is something that can both implement resources and be exposed as resources. For example, the inbox of an agent can be exposed as a resource, with other agents being able to interact with it by sending a POST request to an inbox URL, creating a new message. Alternatively, an agent could expose aspects of its internal state as resources. For example, a bidding agent that participates in an auction could expose its bidding strategy as a resource through a well-defined URL. External systems would then be able to update the strategy by sending a PUT request to that URL, updating the bidding strategy of the agent. In this sense, an AOMS is a community of autonomous agents that expose a related set of resources.

In terms of exposing resources, we adopt the view that each resource is associated with exactly one agent. That said, not every agent needs to be associated with a resource. This offers an interesting dichotomy: an AOMS consists of a set of agents that are mapped to resources and a set of agents that are not. In essence, agents that expose resources form part of the interface of the microservice and must be externally accessible. All other agents are part of the implementation detail and so do not require an external interface. This view offers many parallels with human organisations. For example, in a company, sales and customer support staff are engaged to interact with the public while many other staff are engaged to implement internal functions necessary to run the company. From a software perspective, it also correlates nicely with notion of visibility: externally facing agents can be classed as public agents and all others as private agents.

Building on these notions, we introduce the idea of a class of AOMS that represents reusable components of larger microservices architectures. These AOMS would provide standarised implementations of common auctions [21, 22] and organisational structures [11]. The internal mechanics of these implementations would use agents, but the external interface would be realised as a set of REST resources, allowing both AOMS and POMS to use them. We term this approach **Organisation as a Service (OaaS)**.

3.2 Social Ability and MAMS

Social ability is a key requirement of MAS - it represents the impetus to move away from individual problem solvers towards networks of collaborative/cooperative problem solvers. Traditionally, interaction has been modelled through the use of Agent Communication Languages (ACLs) like FIPA-ACL [19]. More recently, the perceived failures of these ACLs has seen the emergence of new breeds of ACL, such as the Blindingly Simple Protocol Language (BSPL) [42]. Alternatively, other researchers have argued for other modes of interaction based on concepts like shared Artifacts [35], which consider support for interaction to be the responsibility of the environment. This style of model has been discussed in detail in the context of adopting the environment as a first class entity in MAS development [44]. In practice, it has become clear that the adoption of a single approach is often insufficient and many MAS now promote a hybrid communication strategy that combines both approaches. This hybrid view fits well with a microservices approach. As such, we consider a number of complementary modes:

- *Interaction through RESTful APIs.* One potential model of multi-agent interaction discussed in Section 3.1 is to expose the knowledge (state) and capabilities (actions) of an agent directly as resources. Such a model is potentially appealing because it would allow any other component to interact seamlessly and transparently with an AOMS through the associated RESTful API. In such a model, an agent could manage *virtual resources* that are internal abstractions of the agent's internal states or the state of the environment it inhabits. This state is encoded within the agent's beliefs, which are maintained through perception. Changes to the virtual resource through RESTful calls could lead to changes in the agent's beliefs and ultimately drive its actions.
- *Interaction through shared artifacts.* POMS represent resources that are exposed through well-defined interfaces. This is somewhat similar to the notion of an artifact as described in [35]. State updates can be easily achieved through REST operations. It is precisely this model that we envisaged in Section 3.1 where we discussed the idea of a bidding strategy resource being used to allow external systems to update their bidding strategy for an auction.

- *Interaction through high-level communication.* AOMS can represent single or multi-agent systems. Within an AOMS, each externally accessible (public) agent can be uniquely identified by a Uniform Resource Identifier (URI). While we have previously focused on the idea that agents could expose "virtual resources", it is also possible to view the inbox of an agent as a concrete resource, and message passing as being equivalent to a POST operation applied to that resource. Such a model could apply to both locally (the view adopted in this paper) and remotely hosted inboxes [5]. To be clear, this approach is different to the pure RESTful API based communication discussed earlier in that it is more constrained. Specifically, high-level communication refers to communication based on human models of interation [6] (e.g. FIPA-ACL [19]).
- *Interaction through conversations.* As an interesting next step to exposing an agent's inbox as a resource, it is possible to introduce the notion of an agent whose protocols and conversations are exposed as resources. This could be a particularly appealing model as it would allow other agents to understand what protocols an agent knows (possibly including semantic descriptions of those protocols). Additionally, links representing valid responses could be associated with the messages to direct the flow of the conversation. Such a model could build on internal conversation managers like the Agent Conversation Reasoning Engine [30, 31]. For example, if agent A initiates a conversation based on the FIPA Request Protocol with agent B, then A would generate a conversation resource (and associated URI) that encapsulates the conversation. This would make the conversation persistent, referable and lend itself concepts such as accountability and normative behaviour [43].

3.3 Leveraging an Industry Strength Tool Ecology

Microservices has established itself as the leading approach to developing large scale systems in industry [12]. It is recognised as a key tool in managing complexity, maintaining agility and improving the effectiveness of the development team [34]. Key to this success are a suite of tools and components that help to ease the development and deployment process:

- *Containers:* Containerisation sits at the forefront of modern distributed computing, representing a lightweight class of virtualisation platforms. Containers provide a lightweight portable run-time while allowing for development, testing and deployment and the ability to communicate between containers [37]. Key to their success is the focus on single applications/services rather than full Operating Systems coupled with libraries of pre-configured images of services that can be easily downloaded, deployed and tweaked to meet current needs. The dominant implementation of container technology is Docker[1].
- *Container Orchestration:* Effective deployment and management of containerized systems is a key issue. This has led to the emergence of orchestration tools that automate the management and monitoring of large clusters of containers. The dominant player in this space is Kubernetes[2].
- *Infrastructure Services:* The effective deployment of microservices requires more than container orchestration. A range of additional infrastructure services are needed to realise truly reliable deployments. For example, [25] argues that a container orchestrator should facilitate scalability, load balancing, service deployment and discovery, and possibly service migration. [10] adds concepts such as message security, service proxies and data storage infrastructures. For example, when Netflix re-engineered their system to adopt a microservices approach, they were forced to develop a range of technologies that previously were not available [29], including: a load balancer called "Ribbon", and a discovery service called "Eureka".
- *Design Patterns:* Increased adoption of microservices had led to emerging awareness of what works and what does not. Architectural design patterns have been proposed encoding this best practice [10], with the most prominent patterns being: API Gateway, Publish/Subscribe, Circuit Breaker, Proxy and Load Balancer. Many other patterns are discussed online[3]. It is interesting to note that the API Gateway, Pub/Sub and Proxy patterns correspond to KQML facilitator patterns proposed in the early 1990s [18].

4 ILLUSTRATING MAMS

To provide some context for our approach, we have built a prototype based on ASTRA [8, 9], a variation of the AgentSpeak(L) programming language [4, 38]. ASTRA was used because it offers a closer integration with Java than many other AOP languages [8] as it exposes object references to the logical layer of the agent. This allows the direct representation of relations between those objects. It also provides a custom event model (part of its module mechanism) that enables developers to introduce new event types into the language. These features have proven particularly useful in the development of this prototype.

In the creation of this prototype, 4 main aims have been achieved: the provision of a mechanism for implementing a REST interface to virtual resources managed by an agent; the provision of a mechanism that enables agents to interact directly with REST resources; the exposing of the agents inbox as a REST resource; and the illustration of the MAMS concept.

4.1 Instantiating the MAMS model

Before launching into the technical (and ASTRA) specific details of our prototype implementation, we start with some more general refinements that could be applied to any prototype implementation. Specifically, we make the following assumptions:

- As is normal practice, it is assumed that each microservice will run in a separate Java Virtual Machine hosting a single HTTP server. In this way, each microservice instance can be referenced by the hosts name and the port number through which the HTTP server is exposed.

[1] http://docker.com

[2] https://kubernetes.io/

[3] http://microservices.io

- Each public agent will register with and be exposed by this web server. The agent will be treated as a resource and associated with a globally unique URL of the form: `http://<host:port>/<agent-name>`

In our view the agent should not be directly addressable as a REST resource. By this, we mean that you cannot perform HTTP operations on the agent directly. Instead, we argue that an agent is a container for resources and its URL should act as a base for the creation of resource-specific URLs. In our implementation, these will be either the agent's inbox or other virtual resources associated with the agent:

- Each public agent inbox will be exposed through a URI: `http://<host:port>/<agent-name>/inbox`
- Virtual resources associated with an agent will be exposed using URLs based on that agents base URL: `http://<host:port>/<agent-name>/<resource-path>`

A possible future step would be to consider how additional aspects of an agent could be exposed. For example a `/beliefs` URI could return a representation of the agents beliefs. More interestingly, this URI could return the **public** beliefs of the agent allowing an agent to share some of its beliefs with other agents while hiding others. The same could be applied to goals or intentions enabling agents to reason about each others beliefs and activities.

It is worth noting that the URL of an agent is, by definition, globally unique. In our view, this is a simple and elegant solution to the issue of ensuring uniqueness of agent names. The way it is used to expose the agents inbox also contrasts with the approach advocated in the FIPA standards [19], where messages are posted to a URL that represents the agent communication channel (ACC) for the platform which, in turn, is responsible for delivering the message to the agent. This has a downside, in that the agent is potentially less mobile (because its identity is bound to a specific URL), however these problems are well understood on the web, and existing mechanisms such as URL redirection could be used to alleviate such shortcomings.

4.2 Extending ASTRA for AOMS

To illustrate some aspects of our proposed approach, we have adapted the ASTRA programming language to support the creation of virtual resources and FIPA-ACL based communication via RESTful inboxes. This was achieved through the use of ASTRA modules - Java classes that allow the developer to create custom actions, terms, formulae, sensors, and events [?] through annotated methods. Specifically, we developed a `Http` module that provides an interface to a Java Web Server implemented using Netty[4]; a non-blocking IO library.

The Web Server is implemented using the singleton pattern enforcing the 1 server per Java Virtual Machine (JVM) policy. The `Http` module includes two actions to allow creation of the singleton on either a default port (9000) or a user-specified port. An additional action is provided that enables agents to register with the Web Server. This exposes the agent as was defined in Section 4.1.

As was described in the previous section, valid REST resource URLs include the inbox URL or the URL of any virtual resource associated with the public agents. Incoming HTTP requests are handled in one of 2 ways:

- For requests targeted at an agent's inbox, a custom piece of code is executed that transforms the body of the incoming request into a FIPA-ACL message that is passed directly to the relevant agent. This only occurs if the HTTP request is a POST request and a 200 OK response is returned. All other HTTP verbs result in a 403 Forbidden response. To send messages, we implement a custom *ASTRA Message Service*. This is a component of ASTRA that is used to implement bespoke messaging infrastructures. Specifically, this component is used to implement a mechanism that transforms a FIPA-ACL message into a JSON payload that is send as a HTTP POST to a specified URL (another agent's inbox).
- For all other requests, a custom `Http` event is generated. Event types currently exist for only four HTTP verbs: POST, GET, PUT, DELETE. However, it is trivial to extend this to cover all relevant HTTP verbs. The custom `Http` events contain 3 parameters. Two are references to Java objects that represent the incoming request and a context that can be used to generate the HTTP response. The third parameter is a (logical) list of strings that represent the remaining segments of URL associated with the HTTP request that is being handled. These are the segments that come after the part of the URL that equates to the agent's name. We illustrate how custom events are used in Section 3.3.

4.3 Vickrey Auction as a Service

To illustrate our approach, we implement a Vickrey Auction as a Service that can be used by other microservices. A high-level view of the service can be seen in Figure 1. The implementation consists of three types of agent: a Manager agent, Auctioneer agents, and Bidder agents. A single manager is created (with name "manager"). It is responsible for managing two virtual resources:

- `/manager/items`: a resource that represents all items being sold using the service. External services wishing to use the auction service submit POST requests to this URI containing details of items to be sold. Currently, the item model includes a name, a quantity and a reserve price (that must be reached for the item to be sold). The manager agent assigns a unique id to each incoming request. The `/items/{id}` URL can be used to access the item once it has been created.
- `/manager/clients`: a resource that represents all services that are interested in participating in auctions run through the service. Upon registration, a bidder is created for each client microservice. This bidder is the advocate that acts on behalf of the client service. Interaction between the client and the bidder is mediated via a `/wanted` resource through which the client can indicate what items it is interested in and what its bidding strategy should be for each item (e.g. how many items, and how much to pay per item).

To more clearly illustrate how the service works, we will walk through an example sequence of interactions. First we explore how a client microservice (named "MS1") would interact with the auction service. The key steps are illustrated through the arcs labelled 1-5 in Figure 1:

[4] http://netty.io

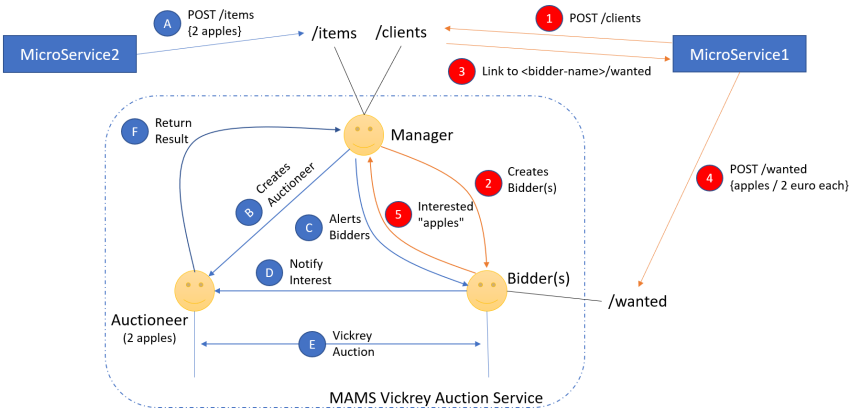


Figure 1: Illustration of the Vickrey Auction as a Service

- **1**: MS1 submits a POST request to `/manager/clients` registering its interest in using the service.
- **2**: The manager verifies MS1, assigns it a unique id and creates a bidder agent matched to that id.
- **3**: The manager responds to the initial POST with a `200 OK` response and the URL of the `<bidder-name>/wanted` resource that has been created specifically for MS1.
- **4**: MS1 submits a post request to the `/wanted/` resource indicating that it wants to purchase 2 apples and is willing to pay 2 euro per apple.
- **5**: MS1's bidder agent registers its interest in auctions of apples with the manager agent.

At this point, an agent has been created to act on behalf of MS1 and it has been assigned the task of buying 2 apples. No further action will take place until another microservice (MS2) decides to sell some apples. To understand the process undertaken by MS2, we will explain the key steps labeled A-F in figure 1:

- **A**: MS2 submits a POST request to `/manager/items` providing details of a type of item it wished to sell (in this case 2 apples). The manager responds immediately with a 200 OK and a link to the item created (through which MS2 can monitor the sales process).
- **B**: At some later point in time, the manager starts the auction process for the item by creating an Auctioneer agent. Tthe delay could be due to a limit on the number of concurrent auctions permitted on the platform, or simply because no other services are currently interested in the item being sold.
- **C**: Once the Auctioneer agent is created, the manager informs all interested bidders that an auction is about to start.
- **D**: Bidder agents decide whether or not to participate. Those that wish to inform the Auctioneer of their interest.
- **E**: The Auctioneer executes a standard Vickrey Auction between all interested bidder agents.
- **F** Details of the winning bid are returned to the manager who updates the item record to reflect the outcome and the winning Microservice is notified of their success.

Figure 2 contains a snippet of ASTRA code that relates to steps B and C above. It implements a simplified solution: the manager agent does not consider whether there are interested agents, but instead creates an auctioneer agent and tells all interested agents about it (if there are no interested agents, then it tells nobody). It executes auctions sequentially. The auctioneer waits 5 seconds to allow interested agents to register their intention to bid, and then starts the auction, in which it waits for a further 10 seconds before assessing all received bids. The highest bid (if one exists) is accepted and the items are sold to the associated microservice.

To deploy our MAMS implementation, it was necessary to make a number of significant changes to the way that ASTRA is compiled and deployed. Traditionally, ASTRA has been supported through a custom Eclipse plugin that integrates with the standard Eclipse build process. This has proven an effective way of allowing developers to write and execute ASTRA programs. However, this build process is not well suited to microservices-based approaches which tend to follow the Continuous Integration/Continuous Delivery (CI/CD) model. To remedy this, we have restructured ASTRA into a set of Maven[5] artifacts. Further, we have re-engineered the ASTRA Compiler to work as part of a Maven Plugin that can be easily integrated into its build lifecycle. Making these changes has allowed us to follow industry best practice to automate the build process and support the automatic creation and deployment of MAMS services. Specifically, we have created a Docker image of our auction system that can be deployed as required.

While the above service is quite simple, we believe that it demonstrates a number of interesting properties:

- We have built and deployed a MAMS that can be used seamlessly by other microservices. This can be easily extended to include external agents through the use of FIPA-ACL and public inboxes.
- By using the MAMS approach, we have implemented a Vickrey Auction that is self-contained and reusable. Other auctions or organisational mechanisms can be implemented in the same way to provide a library of reusable solutions.
- MAMS exposes only the agents/resources that need to interact with external systems. As a result, the Auctioneer, which implements the auction protocol (and decides the winner)

[5] http://maven.apache.org

```
agent Manager {
    ...
    rule $http.post(ChannelHandlerContext ctx,
        FullHttpRequest req, ["items"],
        string bdy) {
      Item item = il.itemFromJson(bdy);
      il.storeItem(item, string id);

      !!auctionItem(id, il.getItemName(item));

      ResponseObject obj = http.createResponse();
      http.setStatus(obj, 200);
      http.setLocation(obj,
          http.myAddress()+"/items/"+id);
      http.sendResponse(ctx, req, obj);
    }

    synchronized rule +!auctionItem(string id,
        string item) {
      !auctioneer("auctioneer"+id, item);
      foreach (interest(string name, item)) {
        send(inform, name,
            available(item, "auctioneer"+id));
      }
    }
    ...
}
```

Figure 2: ASTRA code implementing sequential auctions

is an internal (private) agent. We believe that this offers a more secure solution because it is not possible for agents or microservices (other than the manager and bidders) to interact directly with the Auctioneer.

5 RELATED WORK

Agents and microservices are garnering increased interest within the research community. For example, [26] present a case study in which agents are combined with microservices to implement rule-based eCommerce applications for IoT. The agent aspect of the approach is implemented using EMERALD [27]: a knowledge-based framework built on top of JADE [1].

[28] proposes a collaborative microservices-based model for IoT systems. In their paper, the authors argue, as we do, that microservices can be viewed as being similar to multi-agent systems. However, their approach is more of a design perspective than an implementation strategy and they do not propose the use an concrete agent technologies as we have done in this paper.

[43] offers a view of decentralized multi-agent systems for IoT where agents act as the head of a distributed body consisting of multiple hetereogenous IoT devices. In this paper, sharing of resources (devices) is realised through a normative interaction model that applies positive and negative sanctions in response to interaction. While superficially similar, our approach is really quite different - it does not require agents to act as the head of a service, instead, services and agents are seen as equal partners that are allowed to inter-operate freely as needed by the application. This, when combined with the layered model preferred for REST promotes a view of application that are comprised of a combination of agent based and non-agent based services.

[2] argues for the creation of stateful cloud-native microservices using Akka and Kubernetes. The report clearly highlights the importance of container and container orchestration technologies to modern distributed systems development. This view is one presented by a leading proponent of the Reactive Services/Akka movement. Their clear message is that for technology to be used it industry, it must fit the industry models and views. This paper is our attempt to present that view for MAS technologies.

6 CONCLUSIONS

In this paper, we have argued that adopting a Multi-Agent MicroServices (MAMS) approach offers the potential to deliver truly decentralized multi-agent applications. The specific approach we have used to demonstrate MAMS, described in section 4, is not an attempt to offer a technological solution but instead a vehicle for understanding what a MAMS system may look like and how it may operate. For this reason, we have not attempted to formalization MAMS but instead have followed a more discursive approach.

We conclude by reflecting on the implications of our approach. From a MAS perspective, we believe that some of the main benefits / implications include:

- The embracing of Web models to provide a simple decentralized global unique naming system for agents can act as a basis for driving new decentralized approaches to building MAS. For example, exposing aspects of an agents state and permitting other agents to reason about that state or providing discoverable semantic descriptions that can be used to allow agents to reason about how to interact with one another.
- The decomposition of agent systems into self-contained subsystems that can be independently tested has the potential to improve reliability and promotes the creation of reusable libraries of agents [33].
- The alignment of MAS deployment with industry best practice allows us to leverage the vibrant ecosystem of industry standard tools and components available to microservices based systems and where appropriate adapt them for use with MAS.

From a microservices perspective, we believe that the main benefit is that MAMS enables the seamless integration of agent technology into microservices-based systems. For example, the Organization as a Service (OaaS) model proposed in section 3.1 argues for the creation of shrink-wrapped implementations of various organizational structures that can be used by traditional microservices as mechanisms for mediating access to one another. This can range from implementations of auctions such as the one demonstrated in 4 to fully fledged network or market patterns [11].

ACKNOWLEDGMENTS

This research is funded under the SFI Strategic Partnerships Programme (16/SPP/3296) and is co-funded by Origin Enterprises Plc

REFERENCES

[1] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa. 1999. JADE âĂŞ A FIPA-compliant Agent Framework. *Fourth International Conference on Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM 1999)* (1999), 97–108. https://doi.org/10.1145/375735.376120

[2] Jonas Boner. 2019. How To Build Stateful , Cloud-Native Services With Akka And Kubernetes Tips for running your stateful services as if they are stateless by Jonas Bonér. (2019). https://www.lightbend.com/white-papers-and-reports

[3] Rafael H Bordini, Lars Braubach, Jorge J Gomez-sanz, Gregory O Hare, Alexander Pokahr, and Alessandro Ricci. 2006. A Survey of Programming Languages and Platforms for Multi-Agent Systems. 30 (2006), 33–44.

[4] Rafael H Bordini, Jomi F Hübner, and Renata Vieira. 2005. {Jason} and the Golden Fleece of Agent-Oriented Programming. *Multi-Agent Programming — Languages, Platforms and Applications* 15 (2005), 3–37. https://doi.org/10.1007/0-387-26350-0_1

[5] Jean Paul Calbimonte, Davide Calvaresi, and Michael Schumacher. 2018. Multi-agent interactions on the web through linked data notifications. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-01713-2_4

[6] B Chaib-Draa, A Ken, J Williams, C Hall, and R Kent. 1994. Distributed artificial intelligence: An overview. *Encyclopedia Of Computer Science And Technology* 31 (1994), 215–243.

[7] Andrei Ciortea, Olivier Boissier, and Alessandro Ricci. 2018. Engineering World-Wide Multi-Agent Systems with Hypermedia. In *6th International Workshop on Engineering Multi-Agent Systems (EMAS)*. http://www.fipa.org/repository/standardspecs.html,

[8] R.W. Collier, S. Russell, and D. Lillis. 2015. *Reflecting on agent programming with agentspeak(L)*. Vol. 9387. https://doi.org/10.1007/978-3-319-25524-8_22

[9] A. Dhaon and R. Collier. 2014. Multiple inheritance in AgentSpeak(L)-style programming languages. In *AGERE! 2014 - Proceedings of the 2014 ACM SIGPLAN Workshop on Programming Based on Actors, Agents, and Decentralized Control, Part of SPLASH 2014*. https://doi.org/10.1145/2687357.2687362

[10] Paolo Di Francesco, Ivano Malavolta, and Patricia Lago. 2017. Research on Architecting Microservices: Trends, Focus, and Potential for Industrial Adoption. *Proceedings - 2017 IEEE International Conference on Software Architecture, ICSA 2017* (2017), 21–30. https://doi.org/10.1109/ICSA.2017.24

[11] Virginia Dignum and Frank Dignum. 2001. Modelling Agent Societies: Co-ordination Frameworks and Institutions. *Progress in Artificial Intelligence* 2258 (2001), 7–21. https://doi.org/10.1007/3-540-45329-6_21 arXiv:arXiv:hep-th/0112055v2

[12] Nicola Dragoni, Saverio Giallorenzo, Alberto Lafuente, Manuel Mazzara, Fabrizio Montesi, Ruslan Mustafin, Larisa Safina, Nicola Dragoni, Saverio Giallorenzo, Alberto Lafuente, Manuel Mazzara, Fabrizio Montesi, Manuel Mazzara, and Bertrand Meyer Present. 2017. Microservices : yesterday , today , and tomorrow. *Present and Ulterior Software Engineering* (2017).

[13] Nicola Dragoni, Ivan Lanese, Stephan Thordal Larsen, Manuel Mazzara, Ruslan Mustafin, and Larisa Safina. 2018. Microservices: How to make your application scale. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-74313-4_8 arXiv:1702.07149

[14] Lesser V.R. Durfee, E.H. 1988. Incremental Planning to ControI Time-Constrained Blackboard-Based Problem Solver. *IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS* 24, 5 (1988).

[15] Christoph Fehling. 2015. *Cloud Computing Patterns Identification, Design, and Application*. Technical Report.

[16] Roy T Fielding. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. Dissertation. University of California, Irvine.

[17] Roy T Fielding and Richard N Taylor. 2002. Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology* 2, 2 (2002).

[18] Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire. 1994. KQML as an agent communication language. *Proceedings of the third international conference on Information and knowledge management - CIKM '94* (1994), 456–463. https://doi.org/10.1145/191246.191322

[19] FIPA. 2000. FIPA Standards. http://www.fipa.org

[20] Martin Fowler. 2014. MicroServices: A definition of this new architectural term. https://martinfowler.com/articles/microservices.html

[21] R Guttman and P Maes. 1998. Cooperative Information Agents II, Learning, Mobility and Electronic Commerce for Information Discovery on the Internet, Second International Workshop, CIA' 98, Paris, France, July 4-7, 1998, Proceedings. *Cia* 1435 (1998), 135–147. https://doi.org/10.1007/BFb0053669

[22] Fu Shiung Hsieh. 2006. Analysis of contract net in multi-agent systems. *Automatica* 42, 5 (2006), 733–740. https://doi.org/10.1016/j.automatica.2005.12.002

[23] Muhammad Aslam Jarwar, Sajjad Ali, Muhammad Golam Kibria, Sunil Kumar, and Ilyoung Chong. 2017. Exploiting interoperable microservices in web objects enabled Internet of Things. In *International Conference on Ubiquitous and Future Networks, ICUFN*. https://doi.org/10.1109/ICUFN.2017.7993746

[24] Jeffrey O Kephart and David M Chess. 2003. The vision of autonomic computing. *Computer* 1 (2003), 41–50.

[25] Nane Kratzke. 2014. A lightweight virtualization cluster reference architecture derived from open source paas platforms. *Open Journal of Mobile Computing and Cloud Computing* 1, 2 (2014), 17–30.

[26] Kalliopi Kravari and Nick Bassiliades. 2018. A Rule-Based eCommerce Methodology for the IoT Using Trustworthy Intelligent Agents and Microservices. https://doi.org/10.1007/978-3-319-99906-7_22

[27] Kalliopi Kravari, Efstratios Kontopoulos, and Nick Bassiliades. 2010. EMERALD: A multi-agent system for knowledge-based reasoning interoperability in the semantic web. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6040 LNAI (2010), 173–182. https://doi.org/10.1007/978-3-642-12842-4_21

[28] Petar Krivic, Pavle Skocir, Mario Kusek, and Gordan Jezic. 2018. Microservices as agents in IoT systems. In *Smart Innovation, Systems and Technologies*. https://doi.org/10.1007/978-3-319-59394-4_3 arXiv:arXiv:1607.08131v1

[29] Andrew Leung, Andrew Spyker, and Tim Bozarth. 2017. Titus: Introducing Containers to the Netflix Cloud. *Queue* 15, 5 (2017), 30.

[30] David Lillis. 2012. *Internalising Interaction Protocols as First-Class Programming Elements in Multi Agent Systems*. Ph.D. Dissertation. University College Dublin. https://doi.org/10.13140/RG.2.1.1573.7040/2

[31] David Lillis, Rem W. Collier, and Howell R. Jordan. 2013. Evaluation of a Conversation Management Toolkit for Multi Agent Programming. In *Programming Multi-Agent Systems - 10th International Workshop, ProMAS 2012, Valencia, Spain, June 5, 2012, Revised Selected Papers*, Mehdi Dastani, Jomi F. Hübner, and Brian Logan (Eds.). Vol. 7837. Springer Verlag Heidelberg, 90–107. https://doi.org/10.1007/978-3-642-38700-5_6

[32] David Lillis, Rem W. Collier, Fergus Toolan, and John Dunnion. 2007. Evaluating Communication Strategies in a Multi Agent Information Retrieval System. In *Proceedings of the 5th European Workshop on Multi-Agent Systems (EUMAS '07)*. Hammamet, Tunisia.

[33] Michael Luck, Peter McBurney, and Chris Preist. 2003. *Agent technology: enabling next generation computing (a roadmap for agent based computing)*. AgentLink.

[34] Ekaterina Novoseltseva. 2019. Benefits of Microservices Architecture Implementation. https://dzone.com/articles/benefits-amp-examples-of-microservices-architectur

[35] Andrea Omicini, Alessandro Ricci, Mirko Viroli, Cristiano Castelfranchi, and Luca Tummolini. 2004. Coordination Artifacts: Environment-based Coordination for Intelligent Agents. *International Joint Conference on Autonomous Agents and Multiagent Systems 2004)* (2004), 286–293. https://doi.org/10.1109/AAMAS.2004.10070

[36] Kevin R Page, David C De Roure, and Kirk Martinez. 2011. REST and Linked Data: a match made for domain driven development?. In *Proceedings of the Second International Workshop on RESTful Design*.

[37] Claus Pahl and Brian Lee. 2015. Containers and clusters for edge cloud architectures–A technology review. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*. IEEE, 379–386.

[38] A. Ricci, R.H. Bordini, J.F. Hubner, and R. W. Collier. 2018. AgentSpeak (ER): An Extension of AgentSpeak (L) improving Encapsulation and Reasoning about Goals. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (2018).

[39] Mark Richards. 2015. *Microservices vs. service-oriented architecture*. O'Reilly Media.

[40] S. Russell, H. Jordan, G.M.P. O'Hare, and R.W. Collier. 2011. *Agent factory: A framework for prototyping logic-based AOP languages*. Vol. 6973 LNAI. https://doi.org/10.1007/978-3-642-24603-6_13

[41] Ivan Salvadori, Alexis Huf, Ronaldo dos Santos Mello, and Frank Siqueira. 2016. Publishing linked data through semantic microservices composition. *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services - iiWAS '16* November 2017 (2016), 443–452. https://doi.org/10.1145/3011141.3011155

[42] Munindar P. Singh. 2014. Bliss: Specifying declarative service protocols. In *Proceedings - 2014 IEEE International Conference on Services Computing, SCC 2014*. https://doi.org/10.1109/SCC.2014.39

[43] Munindar P Singh and Amit K Chopra. 2017. The internet of things and multiagent systems: Decentralized intelligence in distributed computing. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1738–1747.

[44] Danny Weyns, Andrea Omicini, and James Odell. 2007. Environment as a first class abstraction in multiagent systems. *Autonomous Agents and Multi-Agent Systems* 14, 1 (2007), 5–30. https://doi.org/10.1007/s10458-006-0012-0

[45] Michael Wooldridge and Jennings Nicholas R. 1995. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review* 10, 2 (1995).

[46] Olaf Zimmermann. 2017. Microservices tenets: Agile approach to service development and deployment. *Computer Science - Research and Development* (2017). https://doi.org/10.1007/s00450-016-0337-0

Digital Companion for Industry

Artificial meets human intelligence

Mareike Kritzler
Siemens Corporate Technology
Artificial & Human Intelligence
Research Group, Berkeley,
California, USA,
mareike.kritzler@siemens.com

Jack Hodges
Siemens Corporate Technology
Artificial & Human Intelligence
Research Group, Berkeley,
California, USA,
jack.hodges@siemens.com

Dan Yu
Siemens Corporate Technology
Artificial & Human Intelligence
Research Group, Berkeley,
California, USA,
kimberly@siemens.com

Kimberly García
Siemens Corporate Technology
Artificial & Human Intelligence
Research Group, Berkeley,
California, USA,
garcia@siemens.com

Hemant Shukla
Siemens Corporate Technology
Artificial & Human Intelligence
Research Group, Berkeley,
California, USA,
hemant.shukla@siemens.com

Florian Michahelles
Siemens Corporate Technology
Artificial & Human Intelligence
Research Group, Berkeley,
California, USA,
florian.michahelles@siemens.com

ABSTRACT

The industrial domain offers a high degree of standardization, a variety of very specialized use cases, and an abundance of resources. These characteristics provide perfect conditions for Digital Companion systems. A Digital Companion is a cognitive agent that assists human users by taking on three roles: as guardians, assistants or mentors, and partners. This paper describes the characteristics, conceptual architecture, use cases and open challenges regarding Digital Companions for industry.

KEYWORDS

Digital Companion, Digital Twin, Cognitive Agent, Semantic Technologies, Human Machine Interaction.

ACM Reference format:

Mareike Kritzler, Jack Hodges, Dan Yu, Kimberly García, Hemant Shukla and Florian Michahelles. 2019. Digital Companion for Industry: Artificial meets human intelligence. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316510

1 Introduction

Digital innovation in industry will result in the creation and collection of massive amounts of digital machine data during a machine's lifetime. New ways of human machine interaction enable proactive machines that evolve from bare tools to smart companions [3]. The challenge is to provide actionable information for human users and to find the right balance of proactivity for machines such that humans and machines can collaborate more effectively and meaningfully together. We propose the development of a Digital Companion, which will digest, integrate, share, and act upon information with humans so that they can focus on meaningful tasks.

A Digital Companion [1] is a helpful cognitive agent that knows the habits of its users, talks to them, entertains them, and can assist them in performing tasks. A Digital Companion considers the implications of empathy, and theory of mind: they learn by imitation and assist in everyday tasks. The term companion has been around for almost two decades now [2]. It builds upon the earlier concept of social robots [4] which describes autonomous robots able to interact and communicate with humans using human-like qualities of interactions, such as trust and empathy. These companions provide useful assistance in a socially acceptable manner. Sociable robots focus on imitating human interaction, and non-verbal communication (e.g., nodding, acknowledgement by posture, and feedback emphasis).

A Digital Companion must have an understanding of machines, components, connectivity, component functionality, component assembly, procedures, and of human-related concepts such as needs/goals/responsibilities, plans, actions, roles, themes, etc. It has to have access to and choose from various forms of presentation and explanation (modality) about a specific problem domain such that it can seamlessly assist users in their work, guide them through new situations, and provide help when and how necessary. Additionally, it has to have the ability to understand its surroundings/context, i.e. terminology, involved objects, users, and roles, and how they relate to the task at hand. A Digital Companion enhances the decision-making capabilities and actions of a person instead of merely taking over. To provide seamless interaction between humans and machines, Digital Companions need to be trustworthy, act for

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316510

the benefit of humans, take care of privacy issues, follow safety guidelines, and above all not be annoying or waste our time in any way.

Popular digital assistants (e.g., Alexa, Google home, Google duplex) are good examples of how humans are getting used to relying on devices to perform simple daily tasks. Those assistants respond to specific requests based on simple pattern matching but do not actually understand what is being said to them. Even with such obvious limitations they have proven quite useful and people are beginning to treat them as though they have greater intelligence and understanding than they do. A Digital Companion, instead, is an entity that understands the situation a human is in, understands the role he is currently playing, his needs, and his goals. A Digital Companion is contextually responsive and able to react to current events.

In this paper we describe the characteristics, a conceptual architecture, industrial use cases, and conclusions on open challenges for developing Digital Companions for industry.

2 Characteristics of a Digital Companion

A Digital Companion consist of three key characteristics: (1) access to shared knowledge (knowledge base), (2) adaptability to users and their context (adaptability), and (3) embodiment as an autonomous entity (embodiment).

2.1 Knowledge Base

In order to enable a Digital Companion to provide useful recommendations and assistance to human users, it needs to have access to a shared knowledge base which describes the properties of entities and the relations between them as well as procedures to transition from one status to another. The underlying foundation of the knowledge base is a set of ontologies, which specify the semantic models underlying data. By integrating different ontologies, different types of data can be related through inference. The ontologies forming the knowledge base can be extended and revised as new data arrives, either explicitly edited or implicitly obtained by observing the user [5]. In a world of agents, it is important that the concepts represented in ontologies be validated by open communities, so it is advantageous to use standardized or standard-compliant ontologies such as SSN, and QUDT whenever possible [10].

Another source of information is established repositories of data and ontologies, such as DBpedia, which provide some generic relationships between entities and terminology even if the ontologies aren't standardized. With the help of machine learning and natural language processing, the companion can extract useful relations from repositories such as these and conceptualize them into the knowledge base. Using these approaches, a Digital Companion can evolve with the ever-changing world (domain) and remain aware of the current context. There is also a need to account for the provenance of data (using, for example, the PROV-O Provenance ontology) such that the knowledge can be assessed by its consumers in terms of credibility and trustworthiness. The knowledge base can answer what it knows, and from where the knowledge was obtained. A Digital Companion represents a helper and mediator that navigates a user through various tasks and seamlessly obtains relevant information by shielding the user from the complexity of underlying systems.

A Digital Companion is also able to extract tacit knowledge from the real world and from the user, e.g. operational knowledge a user exhibits by his actions but would not necessarily be able to express explicitly. When Digital Companions are widely adopted into a particular domain, all the observed data such as best practices, rules of thumb, and common-sense knowledge can be statistically derived and then stored back into a structured information model.

2.2 Adaptability

Another important aspect of Digital Companions is the ability to continuously learn about their users and the associated processes in real-world contexts. Each user has different preferences, backgrounds and experiences. Therefore, Digital Companions have to adapt to the individual user, construct a model about the user, and follow specific preferences and needs. A Digital Companion establishes trust with its user by providing the right information in the right manner at the right time. This requires constructing a model about the users' skills and experiences, history of actions, needs and preferred ways of interaction. A Digital Companion has the means to estimate a user's physiological, emotional, and affective state and builds up a model of the user's level of support and affective responses. A Digital Companion knows how often a user needs to be commended and what forms of appreciations work for them.

A Digital Companion will also adopt the concept of reinforcement learning, which allows users to encourage or discourage certain behaviors using repetition. Users can give feedback to their Digital Companion whenever an action is taken. If an action was repeatedly deemed inappropriate under the same kinds of situation, the companion can learn and avoid performing such an action. In addition to user induced learning, a Digital Companion can learn on its own by observing and collecting data, and then generalized it into the knowledge base. Since a Digital Companion is contextually aware, a user can teach the companion by adding more specific instances of knowledge into the information model. A companion can remember things like preferred route of travel, personal schedule, and instructions to perform specific tasks.

2.3 Embodiment

A Digital Companion can have many embodiments and appearances, which means it resembles a character with a name referring to an entity. This manifestation supports the notion of familiarity and encourages users to interact with a Digital Companion and to engage in dialog with it. The appearances of an embodiment of a Digital Companion range from a simple chatbot to a simulated face in Augmented Reality.

Depending on the domain, the human-computer interface will have to feel natural to its intended users. For example, a factory operator's Digital Companion has access to an information model of a SCADA system based on the OPC UA

standards and related companion models. The Digital Companion can then potentially interoperate with more OPC UA certified devices but also allows subsequent queries into the system. Meaningful actions such as maintenance reminders, anomaly alerts, and troubleshooting support can be generated according to the role or the skillset of the operator. The operator can communicate with the Digital Companion via voice commands that are received, for example, by a wearable wrist watch.

3 Conceptual Architecture

The architecture of a Digital Companion (shown at Figure 1) must acknowledge that it interacts with a human user (at 2) and the world (at **3**). Users interact with the world through their senses. Humans acquire information using sensory perceptors, such as their eyes and ears, and humans affect the world using perceptual effectors, such as speech and touch. A Digital Companion interacts with a user using the same kinds of sensory information, depending on the context. That is, through language, visual, tactile, etc. information. At the same time, a Digital Companion also interacts with the world (at **3**) but is less restricted than a user might be since it can work directly with raw signals and it can interact directly with systems using well-defined APIs. An example of what a Digital Companion can do that a human user cannot, is to go beyond human sense, e.g. sense frequencies as input outside the human perception spectrum between about 20Hz and 20,000Hz.

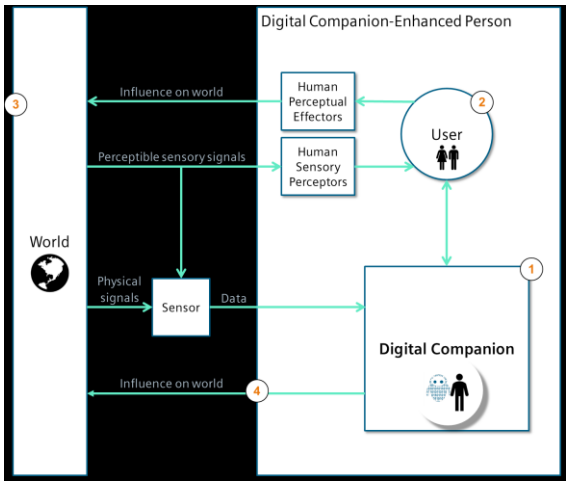


Figure 1: Digital Companion - overall interaction model

The Digital Companion portion (at **1** in Figure 1) is expanded in Figure 2. In this diagram, a user is depicted as a *Human-Machine Interface* (at **2** in Figure 2) which receives and translates perceptual information from a user, such as speech, into a *Conceptual Model* (at **3**), and translates conceptual information back into perceptual information for a user. As mentioned in the discussion of Figure 1, a Digital Companion can interact with the world. In Figure 2 this is shown with a *Data Interface* (at **4**). Also mentioned in Figure 1, the *Data Interface* enables a Digital Companion to interact with many kinds of information and APIs.

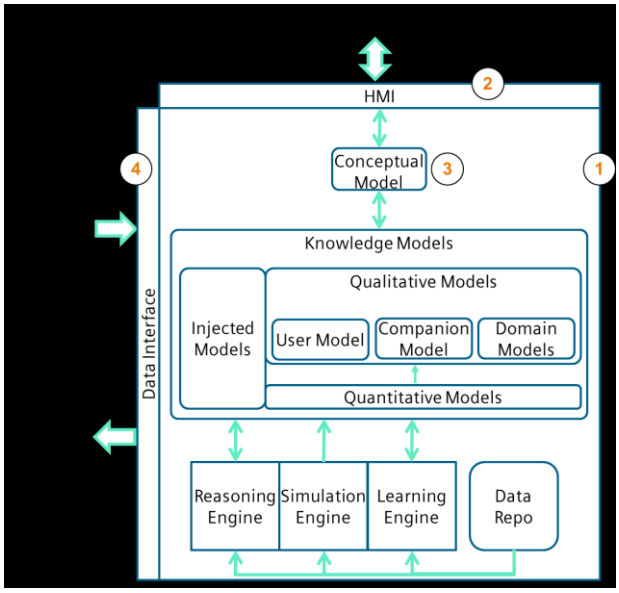


Figure 2: Digital Companion - internal components

Core to a Digital Companion is the ability to operate on conceptual information from a user (by using a Conceptual Model, at **3**) and digital information from the world (at **4**), to understand each type of information, but also to react to it. In order to do so, a Digital Companion must have a model of a user as well as a model of the world, each of these, along with the companion's own model must be malleable, in that they can change over time. These models are shown expanded in Figure 3:

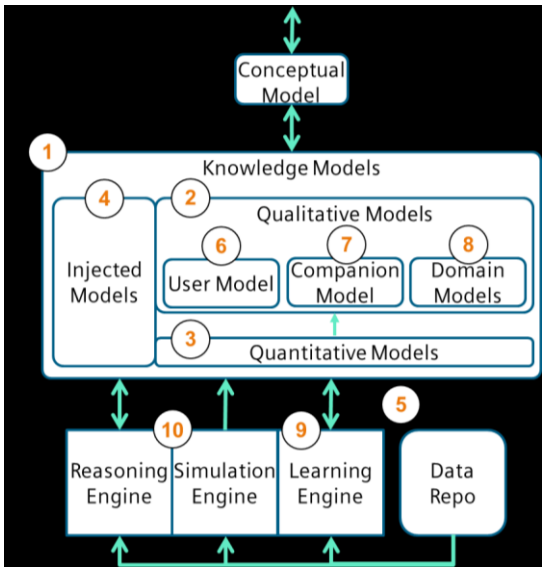


Figure 3: Digital Companion - system architecture

A focal point of the Digital Companion is its knowledge of the world, shown in Figure 3 as an integrated set of *Knowledge Models* (at **1**). These models are further divided into 3 types: *Qualitative Models* (at **2**), *Quantitative Models* (at **3**), and *Injected Models* (at **4**).

Qualitative models are what we think of as conceptual or symbolic models and can be used to represent and integrate any kind of concept. Quantitative models are algorithmic and numerical. Think of a CAD model of geometry or a differential equation as quantitative models. In Figure 3 we show *Quantitative Models* feeding *Qualitative Models* because there can be a mapping from the features and values of *Quantitative Models* to *Conceptual Models*, but not the other way around. *The Injected Models* (again, at **4**) are a specialized form of models used to represent commonsense rules, which differ from the other models in that they are immutable. They can be used but not modified. We use the term injected because they are not learned in this architecture but represent culturally-relevant commonsense knowledge.

Users and Digital Companions each have a *Qualitative Model* that is used by a Digital Companion to make sense of information that becomes available to the system (all incoming data at **5**). Both, the model and the data represent why a user acts the way he does, or why he asks the kinds of questions he asks, which provides a baseline for how to interact with him. This *User Model* (at **6**) must be adaptive, because a Digital Companion will learn how to interact with its user over time, and so the model must change over time as well. Likewise, the *Companion Model* (at **7**) will change over time. The last group in this cluster are the *Domain Models* (at **8**), which represent all the domain knowledge the system has access to. These models are also mutable.

The *Learning Engine* (at **9**) is used to improve all *Qualitative Models*. The *Simulation Engine and Reasoning Engine* (at **10**) are similar but different. They are similar because they are comprised of similar computational mechanisms. They are different because the *Simulation Engine* is used to perform thought experiments and integrate the results, while the *Reasoning Engine* is used to ask and answer questions about known information. The mechanisms associated with decision making, autonomous behavior, model improvement, planning, and problem solving are modules that take model information and data, analyze it, map it into an action model, and act.

4 Roles and Use Cases

This section provides various use cases and matches them to classification of [2] distinguishing between Digital Companions as guardians, assistants or mentors, and partners.

4.1 Digital Companions as Guardians

Performing certain tasks in the workplace does not only require specific skills but also exposes workers to different types of hazards, e.g. drilling requires training and continuous drilling in loud environment can cause hearing loss. Successful task completion requires selecting the right worker and ensuring safe execution. Potential hazards should be mitigated with existing knowledge about safety from recommendation standards. This can be accomplished by combining a semantic reasoner with knowledge models that codify complex workplace safety constraints published by the Occupational Safety and Health Administration (OSHA) while integrating real-time data from sensors (e.g., a visual physiology tracking device) [7]. A mobile system that knows OSHA rules and locations can be used to track personal protective equipment at job sites [6].

Building a Digital Companion based on well-known regulations and standards allows users to trust not only the suggestions of the Digital Companion but also the execution of instructions. This example shows the importance of contextual knowledge as different physical limitations and different workstations require different plans and personal protective equipment.

4.2 Companions as Assistants/Mentors

A Digital Companion can help human users by providing personal assistive services to fulfill tasks, which they themselves would otherwise be unable to perform. With the help of a Digital Companion a user can then develop new skills based on their individual requirements while having a safe guard in place that prevents more severe failures or even harm to the user.

An approach to design scalable and flexible agent-based manufacturing systems that integrates automated planning with multi-agent-oriented programming is presented in [9]. The system uses the Web as an application architecture (and not just as a transport layer), which facilitates the seamless integration of geographically distributed production cells. In a prototypical production cell featuring an industry-grade robot, an augmented reality interface assists the human workers with instructions and guidance following through the manufacturing process.

4.3 Companions as Partners

A Digital Companion can be embodied as an intuitive gateway to complex, distributed information and support during the decision-making process and discuss alternatives. For example, a factory manager needs to make sure that the geographic workforce distribution is sufficient and that the status of factory assets is all in good condition. Therefore, managers need to have easy and intuitive access to the locations of their workforce and to product lifecycle information. To create a partner-like Digital Companion, Semantic Technologies and Augmented Reality are combined to provide context-specific information and allow for hands-free and ubiquitous access to distributed information sources anywhere on a factory floor. This integrated information platform acts as a partner for managers in their decision-making process regarding workforce, assets or tasks [8].

5 Conclusions and Next Steps

This paper introduced the idea of a Digital Companion for industry and outlines its characteristic, architecture and selected use cases. The presented use cases illustrate the various Digital

Companion roles of guardian, assistant, mentor, and partner which extend the capabilities of human users instead of replacing them.

The creation of Digital Companions provides several technical challenges for the research community. Firstly, knowledge and data are the core powering the logic, thus, what kinds of data sources and knowledge bases are suitable for a Digital Companion? Secondly, for a Digital Companion to adapt to new contexts, how can a Digital Companion understand the appropriate context and learn only when and what is meaningful? Thirdly, a Digital Companion interacts with diverse users, how can a Digital Companion strike a balance between being useful and being overwhelming to the users? Lastly, users have to trust that the companion will provide accurate information, keep personal data private, prioritize the users' safety and comfort, and not waste user's time. This trust has to be established and maintained throughout the life cycle of a Digital Companion. How can the Digital Companion evolve within a boundary of trust?

REFERENCES

[1] Peltu, Malcolm and Wilks, Yorick, Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues (January 1, 2008). http://dx.doi.org/10.2139/ssrn.1308500

[2] Böhle, K.; Bopp, K. What a vision: The artificial companion. A piece of vision assessment including an expert survey. Science, Technology & Innovation Studies (STI Studies) 10(2014)1, S. 155-186

[3] Roco M.C. & W.S. Bainbridge, (eds.), 2002. Converging Technologies for Improving Human Performance, NSF-DOC Report, June 2002, US.

[4] Cynthia Breazeal, Toward sociable robots, Robotics and Autonomous Systems, Volume 42, Issues 3-4, 2003, Pages 167-175.

[5] S. Mayer, J. Hodges, D. Yu, M. Kritzler and F. Michahelles: An Open Semantic Framework for the Industrial Internet of Things. In IEEE Intelligent Systems, vol. 32, no. 1, pp. 96-101, Jan.-Feb. 2017.

[6] M. Kritzler, M. Bäckman, A. Tenfält, F. Michahelles: Wearable technology as a solution for workplace safety. MUM 2015: 213-217]

[7] A. Shafei, J. Hodges and S. Mayer: Ensuring Workplace Safety in Goal-based Industrial Manufacturing Systems. In proceedings of Semantics 2018.

[8] I. Mizutani, M. Kritzler, K. García, F. Michahelles: Intuitive interaction with semantics using augmented reality: a case study of workforce management in an industrial setting. IOT 2017: 8:1-8:7

[9] Andrei Ciortea, Simon Mayer, and Florian Michahelles. 2018. Repurposing Manufacturing Lines on the Fly with Multi-agent Systems for the Web of Things. AAMAS '18.

[10] Hodges, J., Garcia, K. & Ray, S. Semantic Development and Integration of Standards for Adoption and Interoperability. *Computer*, 50(11), 26-36. IEEE. 2017.

Towards Enabling Internet-Scale Context-as-a-Service: A Position Paper

Alexandru Sorici
University Politehnica of Bucharest
Bucharest, Romania
alexandru.sorici@cs.pub.ro

Andrei Olaru
University Politehnica of Bucharest
Bucharest, Romania
andrei.olaru@cs.pub.ro

Adina Magda Florea
University Politehnica of Bucharest
Bucharest, Romania
adina.florea@cs.pub.ro

ABSTRACT

Deploying context management systems at a global scale comes with a number of challenges and requirements. We argue that the hypermedia model and the agent-oriented paradigm help achieve the vision of Context-as-a-Service. We categorize challenges according to context processing concerns and use a scenario to exemplify how the proposed architectural principles help overcome the challenges.

KEYWORDS

context management, context-awareness, context-as-a-service, hypermedia, software agents

ACM Reference Format:
Alexandru Sorici, Andrei Olaru, and Adina Magda Florea. 2019. Towards Enabling Internet-Scale Context-as-a-Service: A Position Paper. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316511

1 INTRODUCTION

The observation of an exponential increase of internet-connected sensors and actuators has been a continuous remark for the past decade in the Internet-of-Things (IoT) community, as well as at the level of the everyday consumer of internet-enabled services. While true in terms of the increase of devices, the global-scale exploitation of the wealth of data brought forth by such devices still trends much behind.

Research and industry communities alike have invested in three main areas of focus: industry, society and environment. Of these, the societal focus area is mostly consumer-oriented and, as such, is the one by which the general public can perceive the advancement of the field. While there has been steady progress, the vision held by the ISTAG group back in 2001, in its proposed scenarios [5], where IoT and Ambient Intelligence (AmI) meet, is still far from reach.

In the Maria scenario [5], for example, a single personal assistant, running on a smartphone, handles interactions with transportation, security, communication and smart home systems. The complexity of these interactions comes from the need to continuously and seamlessly switch between different *contexts*.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316511

Context "is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" [1]. Another definition [2] states that context is the *dressing* of a focus, separating the focus of the user or the application from the "dressing" – data which is not vital, but which can improve reasoning on the focus. While these definitions look general, they also offer an explanation for the complexity of deploying the Maria scenario.

Context-aware applications require reasoning procedures, context information representation and support for dynamic and heterogeneous context providers and consumers, all of which can vary greatly across domains. For instance, in a smart crop monitoring application, context refers to the data retrieved from *static* sensors installed in the field, for example monitoring soil and weather conditions, as well as meta-data about the sensors themselves (battery power, physical location, distance from other sensors). On the other hand, in a smart city application, user focus may change rapidly from the current activity at home or at the workplace, to that of monitoring personal health parameters while executing a physical workout activity. The consumer of context is mobile and changes focus quickly, while the means to derive the required high-level context information may involve data-driven algorithms (machine-learned data), along-side knowledge-driven heuristics.

Apart from the variety of context-aware application domains themselves, the problem of interest for *specific* context information within a domain, or combining information from several domains gave rise to the Sensing-as-a-Service model for IoT [13].

The model tackles the issue of siloed systems, proposing the existence of several stakeholders: the *sensor owner* (e.g. an individual person, a state or a private organization), the *sensor publishers* (organizations implementing means for sensor data collection), the *extended service providers* (organizations that bring added value by analyzing and aggregating sensor data) and the *final consumer*.

The sensing-as-a-service vision carries with it implicit non-functional requirements for *openness* and *scalability*, since the model allows for and encourages organizations to play several of the mentioned roles at once, thus leading to the development of *context prosumers* (producer *and* consumer at the same time). Consequently, context management plays a central role in this vision, precisely because it implies that consumers, with *different interests*, require data from different providers (with their own interests), processed in *custom* ways.

From a context management perspective, implementing the sensing-as-a-service model, as well as the ISTAG vision for AmI, brings about a number of challenges, which we group along the following concerns.

Representation and Reasoning. At global scale, a single context model or reasoning/inference technology cannot be expected for all application domains.Mechanisms that facilitate conversion from one model to another, as well as on-demand execution of various context processing procedures have to be considered.

Context Provisioning. This perspective focuses on how context information is supplied to services that can add value to raw data through inference. The challenge in AmI and sensing-as-a-service lies in the fact that the input base can no longer be considered static. Rather, a *long-lived* means to *interconnect*, *search for* and *select* the most appropriate/available context sources and context processors required for a consumer request is needed.

Query and Dissemination. The large-scale nature of sensing-as-a-service and pervasive AmI means that there must be a means to handle queries/subscriptions with a large selection base. Handling of frequently overlapping queries, as well as result caching should be enabled where appropriate. Another challenge comes from the principle that context information should be consumed as *close* as possible to its production source, thereby increasing the chance of relevance. An organizational scheme is needed that facilitates *local* consumption of context information, while at the same time enabling a *structured dissemination* of context information for remote consumers.

Context Service Deployment. The challenge from the deployment perspective is to find the means to facilitate scalability, search and discoverability of context management services. An organizational scheme that facilitates a *single entry-point* for all context life cycle entities (producers, processors, consumers) is required. The deployment structure must address mobility of context consumers and, more importantly, the *shift* in their focus, which brings a change in the necessary context information.

While these challenges have been addressed individually in existing work, we argue that visions such as sensing-as-a-service and ISTAG-level AmI are still not realized because of two main short-comings: **(i)** the lack of support architecture and information structuring mechanisms in existing context middleware that would allow for organization of context such that it becomes discoverable / searchable at a large scale, by highly mobile consumers; **(ii)** insufficient support in existing context middleware to *negotiate* between different (possibly conflicting) stakeholder (e.g. context producer, processor, consumer) interests, when considering the sensing-as-a-service model.

To answer to these challenges, we argue that context management middleware must adopt an architectural style that promotes long-lived deployment, openness and evolvability. Specifically, we posit that the fundamental engineering techniques that sustain the Web (e.g. hypermedia-driven interactions, RESTful [6] protocols, knowledge graphs, publish/subscribe mechanisms) can be coupled with the organizational and behavioral principles stemming from the Multi-Agent System (MAS) literature to obtain architectural specifications for Context-as-a-Service (CaaS) systems supporting application areas as large-scale as AmI and Sensing-as-a-Service. While a hypermedia environment facilitates large scale discoverability/searchability, the agent-oriented paradigm allows conceptualization of the sensing-as-a-service stakeholders as autonomous entities upholding policy driven goals.

2 RELATED WORK

Recent survey papers [10, 12] perform a good review of several of the existing context management middleware. While each solution covers many aspects that are essential for the implementation of a context management life cycle, most of the existing proposals either focus on specializing for a particular application domain (e.g. SeCoMan [3] – location, CoCaMAAL [7] – eHealth), or on the engineering effort for one or more of the focus requirements outlined in the introduction. The CoaaS platform introduced in [8], for example, introduces a detailed view of architectural elements responding to challenges of on-demand reasoning and large scale query management (including caching and request prediction). However, the systems proposes a centralized, cloud-based deployment and does not address the issue of organizing context information and searchability/discoverability thereof, beyond a directory-based mechanism.

CONSERT [14] is a context middleware that uses the multi-agent programming paradigm to design autonomous management for each of the context life cycle entities. It further proposes a deployment mechanism that follows an explicit organizational scheme, which uses *Context Dimensions* (well-defined focus points of a consumer that dominate the rest of the perceived context information - e.g. location, activity, role in an organization) to connect the different instances of context agents. While conceptually it answers to the challenges enumerated under the query/dissemination and service deployment categories, it does not meet the requirements for context provisioning and multi-modal, on-demand reasoning capabilities. Furthermore, the middleware has not undergone any real-world scenario validation.

FIWARE[1] is an open-source cloud platform aiming to provide reliable means for developing open, collaborative and mature ecosystems of smart, context-aware, internet-scale applications. It comes close to the set of requirements outlined in the introduction.

The context management reference architecture of FIWARE[2] is based on the interaction between key Generic Enablers (GE) that facilitate the development of a context processing pipeline. The central element is the *Context Broker* GE, which defines the context model and allows for context information storage, update and look-up. The *Context Broker* can be connected to a series of other GEs which complement its functionality (e.g. a *IoT Broker* as a wrapper over IoT devices, *IoT Discovery* to facilitate discovery of device capabilities, *Complex Event Processing* and *Big Data Analysis* GEs for short- and long-term reasoning). A core feature of the FIWARE platform is its reliance on web-based standards for enabling the communication between GEs. The Open Mobile Alliance (OMA) NGSI-9 and NGSI-10 [9] are two specifications for RESTful interaction protocols that define means to semantically describe the capabilities of IoT devices (NGSI-9) and the exchange of context content itself (e.g. publishing, updating, querying – through NGSI-10).

However, while the structural elements are there, the FIWARE platform proposes no conceptual means of organizing the connection of its GEs, so as to enable large-scale, automated discoverability

[1]FIWARE (https://www.fiware.org/, https://www.fiware.org/developers/catalogue/)
[2]FIWARE Context Management Architecture (https://goo.gl/Yvwv5A)

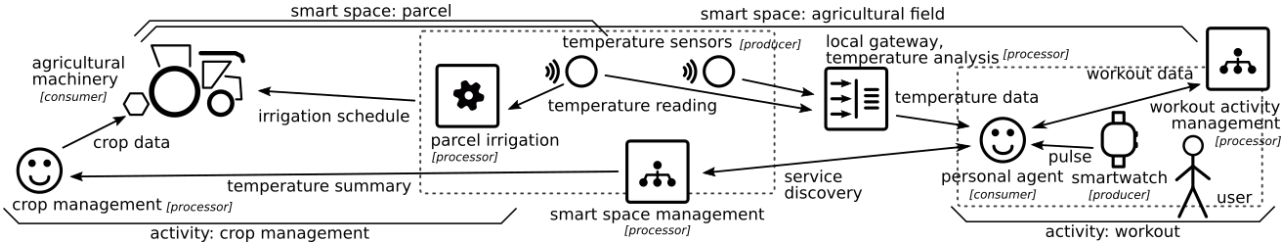


Figure 1: An example of context management services enabling a scenario in which temperature data is used both for crop management and for a user's context-aware activity tracking. Context life cycle entity roles are written in parentheses for each actor. Dotted boundaries denote logical partition of context life cycle entity deployment into *Context Domains* (e.g. parcel, workout), based on spatial and activity *Context Dimensions*.

and dissemination of information. The *Context Broker* interconnection has to be manually specified and is predefined at development time. Furthermore, the existing *Context Broker* implementation does not handle complex (composite) query execution, nor can the *Complext Event Processing* GE manage on-demand context reasoning.

Nonetheless, the takeaway point of FIWARE is that the defined RESTful interfaces contain in their specification the potential to enable all the outlined types of interaction, as required. Either upgrading the existing GEs, or delivering fresh implementations compatible with the NGSI-9 and NGSI-10 specifications has the potential to overcome the challenges.

3 HYPERMEDIA-DRIVEN CAAS

While the exact mechanisms and methods for context management are highly application-dependent, the multitude of entities and stakeholders in a global-scale system for the management of context information requires a suitable underlying architecture, which can answer to the outlined challenges and also fulfill the non-functional requirements of openness and scalability. The purpose of this architectural specification is to lead to the development of systems that support a large-scale *Context-as-a-Service* (CaaS) view, where context management can be offloaded to a network of *context life cycle entities*, each working under its own policy on context production, reasoning, or query management/dissemination.

Such an underlying architecture is the *resource-oriented* model of the Web and the *hypermedia-driven interactions* it supports. Moreover, the *agent-oriented paradigm* can contribute to modeling individual entities in the CaaS ecosystem, by focusing on a perspective centered on individual participants, rather than on the system as a whole.

Hypermedia underpins the World Wide Web as a network of uniquely identifiable (by means of URIs) *resources* interconnected through *web services*. However, to truly exploit the resource-oriented nature for the modeling of context processing entities, their *state* (e.g. context production capabilities, query subscription results, active reasoning mechanisms) has to be explicitly (semantically) described and the means to *change their state* should be clearly identifiable.

In terms of the **context service deployment** challenges, the use of a hypermedia environment and RESTful [6] interactions is not sufficient on its own to facilitate discoverability and efficient search / query propagation. In the architectural specification we envision, context life cycle entities must have a structured means by which to determine their *required* connections. We draw inspiration from the vision of Socio-Technical Networks (STNs) [4] and the proposed deployment organization scheme of CONSERT [14]. All URIs referring to connections between entities of an STN are *typed relations*, meaning that there is a clear semantics attached to them (e.g. ownership, membership, collocation). In CONSERT, *Context Dimensions* and *Context Domains* are concepts defining key *properties* and *values* that relate a context consumer to context sources that might be of interest to his current *focus*. For instance, the focus of the user may be included in various *spatial Context Domains*, also a hierarchy of *activity*-related *Domains*, and also in some *social Domains*. This hierarchical organization of context entities in the same *Dimension* facilitates their management and indicates clearly the context sub-model of which they are responsible.

Let us take the example in Figure 1. A person is taking a jog (workout) in the vicinity of a farm. Two dimensions of context are distinguishable: a *spatial* one and an *activity*-based one. From the spatial perspective, the ⟨`smart space context manager`⟩ keeps track of a hierarchy of two domains – a particular *parcel*, which is ⟨`part-of`⟩ an *agricultural field*. In the context model, an activity has a place where it is carried out, a relation which is modelled semantically and explicitly by the ⟨`personal agent`⟩ (which acts as a *context consumer*) on behalf of the jogger. By this relation, the ⟨`personal agent`⟩ uses a discovery mechanism through which it looks explicitly for a *context processor* responsible for the space in which the jogging activity takes place. The ⟨`personal agent`⟩ then launches a query for the current surface temperature. The result of this query is forwarded to the ⟨`workout activity manager`⟩, where this *context processor* uses it, together with data from the smart watch, to warn the user of excess dehydration, depending on the outside weather conditions and the calories he has burned.

The example illustrates both the model of sensing-as-a-service, since various information is gathered from various sources in the current context; and the model of structured search for context information, since entities are communicating with other entities in the same *Context Domain*.

The critical observation regarding the *deployment perspective* in the given example is that the ⟨`personal agent`⟩ *is* or *becomes* aware of two *Context Domains* (the jogging activity and the agricultural field location) and is able to connect to the relevant *context processors* to retrieve context that "dresses" the current focus of the user (his

well-being while jogging). This is made possible by the explicit, *typed* spatial inclusion and current activity relations which relate the ⟨personal agent⟩ with the ⟨workout activity manager⟩ / ⟨smart space manager⟩.

From a **query / dissemination perspective**, a *Context Dimension*-based organization of *context processors* facilitates query routing. In our example, a manner of handling queries similar to the one described by Mayer et al. [11] is proposed. Both the agricultural machinery and the personal agent subscribe for temperature updates to the ⟨temperature analysis⟩ *processor*. Being mobile consumers, they will request temperature updates for given locations (parcels) within the field. If temperature analysis is managed by different *processors*, queries sent to the processor for the entire field must be routed to the one corresponding to current consumer location. The *Context Dimension* / *Context Domain* based organization of the *processors* enables the same type of query routing mechanism as described in [11] (including *Context Domain*-based range queries).

An *agent-oriented view* of the context life cycle elements in our example is of relevance from the **context provisioning perspective**. Software agents are entities that are able to act autonomously in their environment in order to achieve or maintain their goals.

The owner of the agricultural field is the one that deployed the temperature sensors and the ⟨temperature analysis⟩ *processor(s)*. The ⟨personal agent⟩ may be able to discover the existence of the *processor* but this may not guarantee that he has the *right* to access their information. This use case can be viewed as a sensing-as-a-service instance, where the ⟨temperature analysis⟩ is an *extended service provider* and the ⟨personal agent⟩ a *final consumer*. If we model/design the life cycle entities as *agents*, then access to the temperature data can be subject to an automated, agent based negotiation. The ⟨temperature analysis⟩ *processor* can make use of goal-based policies, that require it to, for example, accept query requests only for a specific user activity context, such as the workout one.

The **representation and reasoning perspective** benefits in two ways from the proposed principles. Context models are application dependent, but the desire for searchability requires the use of a common-ground language. Standardization efforts of the W3C such as the SSN ontology[3] or the Web-of-Things Thing Description[4] can be employed in this case. The agent-oriented view, on the other hand, is suitable for multi-modal and on-demand reasoning. In our example, default reasoning carried out by the ⟨workout activity⟩ *processor* is complemented (upon discovery of temperature data) by a request from the ⟨personal agent⟩. The ⟨workout activity⟩ is asked to execute a knowledge-driven heuristic that combines temperature *and* smart watch context to gauge the user's well-being.

4 CONCLUSIONS

We argue that context servicing at the scale of pervasive AmI settings and sensing-as-a-service applications is lacking because of a missing evolvable and searchable middleware ecosystem, where organized context information enables its effective exploration. We further assess that an agent-based view of the context life cycle entities can appropriately model the required level of autonomy and goal-driven behavior of stakeholders in a sensing-as-a-service setting.

We envision a large-scale CaaS deployment as a conglomerate of computing entities obtaining producing, processing, and consuming context. The multitude and diversity of entities inevitably leads to a highly *heterogeneous* system, that needs to be *loosely coupled*. Not only does the hypermedia model naturally support openness and scalability, but it also answers well to the challenges of CaaS deployment. We also argue that context management systems would benefit from an agent-oriented approach to the implementation of their individual components. What we can take from the MAS domain and use in CaaS is not necessarily how entities can be implemented as agents, but the *agent-oriented perspective*. Moving from a system-wide view to an entity-centered helps designing components that are able to work in highly heterogeneous systems.

[3]https://www.w3.org/TR/vocab-ssn/

[4]https://w3c.github.io/wot-thing-description/

ACKNOWLEDGMENTS

This research was funded by grant PN-III-P1-1.2-PCCDI-2017-0734.

REFERENCES

[1] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*. Springer, 304–307.

[2] Juliette Brézillon and Patrick Brézillon. 2007. Context Modeling: Context as a Dressing of a Focus. In *Modeling and Using Context*, Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, and Laure Vieu (Eds.). Lecture Notes in Computer Science, Vol. 4635. Springer Berlin Heidelberg, 136–149.

[3] Alberto Huertas Celdrán, Félix J García Clemente, Manuel Gil Pérez, and Gregorio Martínez Pérez. 2016. SeCoMan: A Semantic-Aware Policy Framework for Developing Privacy-Preserving and Context-Aware Smart Applications. *IEEE Systems Journal* 10, 3 (2016), 1111–1124.

[4] Andrei Ciortea, Antoine Zimmermann, Olivier Boissier, and Adina Magda Florea. 2016. Hypermedia-driven Socio-technical Networks for Goal-driven Discovery in the Web of Things. In *Proceedings of the Seventh International Workshop on the Web of Things*. ACM, 25–30.

[5] K. Ducatel, M. Bogdanowicz, F. Scapolo, J. Leijten, and J.C. Burgelman. 2001. *Scenarios for ambient intelligence in 2010.* Technical Report. Office for Official Publications of the European Communities.

[6] Roy T Fielding and Richard N Taylor. 2000. *Architectural styles and the design of network-based software architectures*. Vol. 7. University of California, Irvine Irvine, USA.

[7] Abdur Forkan, Ibrahim Khalil, and Zahir Tari. 2014. CoCaMAAL: A cloud-oriented context-aware middleware in ambient assisted living. *Future Generation Computer Systems* 35 (2014), 114–127.

[8] Alireza Hassani, Alexey Medvedev, Pari Delir Haghighi, Sea Ling, Maria Indrawan-Santiago, Arkady Zaslavsky, and Prem Prakash Jayaraman. 2018. Context-as-a-Service Platform: Exchange and Share Context in an IoT Ecosystem. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 385–390.

[9] Srdjan Krco, Boris Pokric, and Francois Carrez. 2014. Designing IoT architecture(s): A European perspective. In *Internet of Things (WF-IoT), 2014 IEEE World Forum on*. IEEE, 79–84.

[10] Xin Li, Martina Eckert, José-Fernán Martinez, and Gregorio Rubio. 2015. Context aware middleware architectures: survey and challenges. *Sensors* 15, 8 (2015), 20570–20607.

[11] Simon Mayer, Dominique Guinard, and Vlad Trifa. 2012. Searching in a web-based infrastructure for smart things. In *Internet of Things (IOT), 2012 3rd International Conference on the*. IEEE, 119–126.

[12] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2014. Context aware computing for the internet of things: A survey. *IEEE communications surveys & tutorials* 16, 1 (2014), 414–454.

[13] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2014. Sensing as a service model for smart cities supported by internet of things. *Transactions on Emerging Telecommunications Technologies* 25, 1 (2014), 81–93.

[14] Alexandru Sorici, Gauthier Picard, Olivier Boissier, and Adina Florea. 2015. Multi-agent based flexible deployment of context management in ambient intelligence applications. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Springer, 225–239.

The First International Workshop on Knowledge Graph Technology and Applications Chairs' Welcome

It is our great pleasure to welcome you to the *WWW 2019 The First International Workshop on Knowledge Graph Technology and Applications*. We received 18 submissions from authors coming from Asia, Europe, North America, and South America. Based on the results of reviews, we finally accepted 8 submissions as full research articles.

The morning session includes two invited talks, one panel and four paper presentations:

- Invited Talks by Denny Vrandecic (Google) and Jure Leskovec (Stanford)
- Panel: Computational Methods about Knowledge Graph by Jie Tang (Tsinghua Unviersity), Jure Leskovec (Stanford), Kuansan Wang (Microsoft), Deborah McGuinness (RPI), and Hassan Sawaf (Amazon)
- Four paper presentations: 1) Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolution Network; 2) How new is the (RDF) news? Assessing knowledge graph completeness over news feed entities; 3) Content based News Recommendation via Shortest Entity Distance over Knowledge Graphs; 4) Trust and Privacy in Knowledge Graphs.

The morning session includes two invited talks, one panel and four paper presentations:

- Invited Talks by Yuqing Gao (Microsoft) and Hassan Sawaf (Amazon)
- Panel: Knowledge Graph Industry Applications by Joshua Shinavier (Uber), Kim Branson (Genentech), Wei Zhang (Alibaba), Shima Dastgheib (numedii), Yuqing Gao (Microsoft), Bogdan Arsintescu (LinkedIn), Fatma Özcan (IBM), and Edgar Meij (Bloomberg).
- Four paper presentations: 1) Building a Knowledge Graph for the Air Traffic Management Community; 2) WebProtege: A Cloud-Based Ontology Editor; 3) Scalable Knowledge Graph construction over text using Deep Learning based Predicate Mapping; 4) An Attention-based Model for Joint Extraction of Entities and Relations with Implicit Entity Features

Ying Ding
Organizing Committee
Indiana University

Chenyan Xiong
Organizing Committee
Microsoft

Raghava Mutharaju
Organizing Committee
Indraprastha Institute of Information Technology, Delhi (IIIT-D)

The First International Workshop on Knowledge Graph Technology and Applications Organization

Workshop Chairs:
Huajun Chen,Zhejiang University, China
Laura Dietz, University of New Hampshire, USA
Ying Ding, Indiana University Bloomington, USA
Wendy Hall, University of Southampton, UK
James Hendler, Rensselaer Polytechnic Institute, USA
Deborah McGuinness, Rensselaer Polytechnic Institute, USA
Edgar Meij, Bloomberg LP, USA
Sam Molyneux, Chan Zuckerberg Initiative, USA
Varish Mulwad, GE Global Research Center, USA
Raghava Mutharaju, IIIT-Delhi, India
Jeff Z. Pan, University of Aberdeen, UK
Xiang Ren, University of Southern California, USA
Jie Tang, Tsinghua University, China
Alex Wade, Chan Zuckerberg Initiative, USA
Mengting Wan, University of California, San Diego (UCSD), USA
Chenyan Xiong, Carnegie Mellon University, USA
Min Zhang, Tsinghua University, China

Organization Committee:
Ying Ding: dingying@indiana.edu
Raghava Mutharaju: raghava.mutharaju@iiitd.ac.in
Chenyan Xiong: chenyan.xiong@microsoft.com

Program Committee:
Soren Auer (auer@tib.eu), Leibniz University of Hannover, Germany
Juan Sequeda (juanfederico@gmail.com), Capsenta, USA
Freddy Lecue (freddy.lecue@gmail.com), CortAIx@Thales, Canada
Steve Gustafson (steven.gustafson@maana.io), Maana, USA
Craig Knoblock (knoblock@isi.edu), University of Southern California, USA
Pascal Hitzler (pascal.hitzler@wright.edu), Wright State University, USA
Tim Finin (finin@cs.umbc.edu), University of Maryland, Baltimore County, USA
Axel Polleres (axel.polleres@wu.ac.at), WU Vienna, Austria
Amelie Gyrard (Amelie@knoesis.org), Wright State University, USA
Pankesh Patel (pankeshlinux@gmail.com), Fraunhofer, USA
Peter Haase (ph@metaphacts.com), metaphacts GmbH, Germany
Muhammad Intizar Ali (ali.intizar@insight-centre.org), Insight Center, Ireland
Krzysztof Janowicz (jano@geog.ucsb.edu), UCSB, USA
Paul Groth (p.groth@uva.nl), University of Amsterdam, Netherlands

Invited Speakers – Morning Session

The First International Workshop on Knowledge Graph Technology and Applications

Denny Vrandecic, Google

Title: To be Announced

Abstract: coming soon

Bio: Denny works at the Google Knowledge Graph. He previously has worked at the Karlsruhe Institute of Technology (2004-2012), the University of Southern California (2010), and as the project director of Wikidata at Wikimedia Deutschland (2012/13). His research interests are massive collaborative systems, knowledge bases, and the Semantic Web.

Jure Leskovec, Stanford

Title: To be Announced

Abstract: coming soon

Bio: Associate Professor of Computer Science at Stanford University. My research focuses on mining and modeling large social and information networks, their evolution, and diffusion of information and influence over them. Problems I investigate are motivated by large scale data, the web and on-line media.

Invited Speakers – Afternoon Session

The First International Workshop on Knowledge Graph Technology and Applications

Yuqing Gao, Microsoft

Title: To be Announced

Abstract: coming soon

Bio: An accomplished technical scientist, innovator and R&D leader in cutting-edge technology research and product development. Proven track record of success (20+ years of successful professional career in leading R&D organizations) in leading rapid technological advancement, innovation, and highly competitive environments. Broad range of skills from initiating research breakthroughs to achieving marketable product development. Renowned expert and technological visionary in the fields of enterprise middleware, cloud computing, data centric computing, workload optimized systems and appliances, business analytics, big data, social media and multimedia, speech and natural language processing.

Hassan Sawaf, Amazon

Title: To be Announced

Abstract: coming soon

Bio: Director of Artificial Intelligence at Amazon Web Services, passionate about opening up new markets and opportunities with smart application of Artificial Intelligence and application-driven research in AI, in particular in language technology, speech processing, computer vision and computational reasoning.

Panel: Knowledge Graph Industry Applications

The First International Workshop on Knowledge Graph Technology and Applications

ABSTRACT

This panel will focus on industry applications related to knowledge graph and showcase how knowledge graph transforms the conventional and unconventional industries to the new era of AI, ranging from innovations in medicine and healthcare, literature search, e-commerce, professional connections, to getting a ride.

Panelists are:

- Joshua Shinavier, Uber

Senior Software Engineer/Research Scientist at Uber, Co-founder of Tinkerpop, specialized on real-time semantics, RDF streams and graph databases

- Kim Branson, Genentech

Head of AI at Genentech, passionate about modeling, and currently developing a general medical inference engine that can be applied to a wide variety of areas, from point of care decision support, triage, insurance risk managment to name a few

- Wei Zhang, Alibaba

A senior staff engineer/director at Business Platform Unit, Alibaba, leading product knowledge graph (PKG) team and Business Platform AI team. He and his team have built a huge PKG with 10 billion of entities.

- Shima Dastgheib, numedii

Data Scientist at Numedii, previously postdoctoral research fellow at Stanford University School of Medicine, developing novel methods to integrate and explore a broad set of biological and clinical data for scientific reproducibility and biomedical discovery.

- Yuqing Gao, Microsoft

An accomplished technical scientist, innovator and R&D leader in cutting-edge technology research and product development. Proven track record of success (20+ years of successful professional career in leading R&D organizations) in leading rapid technological advancement, innovation, and highly competitive environments. Broad range of skills from initiating research breakthroughs to achieving marketable product development. Renowned expert and technological visionary in the fields of enterprise middleware, cloud computing, data centric computing, workload optimized systems and appliances, business analytics, big data, social media and multimedia, speech and natural language processing.

- Bogdan Arsintescu, LinkedIn

Extended technical leadership in systems design for LinkedIn Economic Graph, Google Search and Google Research. Breadth and depth of expertise in building data systems and platforms. Technical leadership and management of software development in both start-up and large companies.

- Fatma Özcan, IBM Almaden

A principal research staff member, and a senior manager at IBM Almaden Research Center. I manage the information management department, working on HTAP (hybrid transactional and analytical processing) systems, large scale machine learning, and natural language querying of data.

- Edgar Meij, Bloomberg

A team lead and senior scientist at Bloomberg. He holds a PhD in computer science from the University of Amsterdam and has an extensive track record in artificial intelligence, information retrieval, knowledge graphs, natural language processing, and machine learning. Before joining Bloomberg he worked at Yahoo Labs on semantic search at web scale using the Yahoo Knowledge Graph. At Bloomberg he leads the team that is responsible for leveraging knowledge graph technology to drive advanced financial insights.

ACM Reference format:

2019. Panel: Knowledge Graph Industry Applications: The First International Workshop on Knowledge Graph Technology and Applications. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317711

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6675-5/19/05
https://doi.org/10.1145/3308560.3317711

Panel: Computational Methods about Knowledge Graph

The First International Workshop on Knowledge Graph Technology and Applications

ABSTRACT

This panel will focus on the cutting-edge computation methods, which can be applied to knowledge graph, such as latest NLP technologies to extract entities and relationships to build knowledge graphs, machine learning or deep learning methods on mining knowledge graph, and intelligent search or recommendations powered by knowledge graphs.

Panelists are:

- Jie Tang, Tsinghua University

Professor and the Vice Chair of the Department of Computer Science and Technology of Tsinghua University. I obtained my Ph.D. in DCST of Tsinghua University in 2006. My research interests include artificial intelligence, data mining, social networks, machine learning and knowledge graph, with an emphasis on designing new algorithms for mining social and knowledge networks.

- Jure Leskovec, Stanford

Associate Professor of Computer Science at Stanford University. My research focuses on mining and modeling large social and information networks, their evolution, and diffusion of information and influence over them. Problems I investigate are motivated by large scale data, the web and on-line media

- Kuansan Wang, Microsoft

Managing Director of MSR Outreach, an organization with the mission to serve the research community. In addition to applying the intelligent technologies to make Bing and Cortana smarter in gathering and serving academic knowledge, we are also starting an experimental website, academic.microsoft.com (powered by Academic API), and mobile apps dedicated to exploring new service scenarios for active researchers like myself

- Deborah McGuinness, RPI

A leading expert in knowledge representation and reasoning languages and systems and has worked in ontology creation and evolution environments for over 20 years. Most recently, McGuinness is best known for her leadership role in semantic web research, and for her work on explanation, trust, and applications of semantic web technology, particularly for scientific applications.

- Hassan Sawaf, Amazon

Director of Artificial Intelligence at Amazon Web Services, passionate about opening up new markets and opportunities with smart application of Artificial Intelligence and application-driven research in AI, in particular in language technology, speech processing, computer vision and computational reasoning.

ACM Reference format:

2019. Panel: Computational Methods about Knowledge Graph: The First International Workshop on Knowledge Graph Technology and Applications. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317712

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6675-5/19/05
https://doi.org/10.1145/3308560.3317712

Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolutional Network

Shumin Deng
Zhejiang University, AZFT Joint Lab for Knowledge Engine
China
231sm@zju.edu.cn

Ningyu Zhang
Alibaba Group
China
ningyu.zny@alibaba-inc.com

Wen Zhang
Zhejiang University, AZFT Joint Lab for Knowledge Engine
China
wenzhang2015@zju.edu.cn

Jiaoyan Chen
University of Oxford
United Kingdom
jiaoyan.chen@cs.ox.ac.uk

Jeff Z. Pan
University of Aberdeen
United Kingdom
jeff.z.pan@abdn.ac.uk

Huajun Chen
Zhejiang University, AZFT Joint Lab for Knowledge Engine
China
huajunsir@zju.edu.cn

ABSTRACT

Deep neural networks have achieved promising results in stock trend prediction. However, most of these models have two common drawbacks, including (i) current methods are not sensitive enough to abrupt changes of stock trend, and (ii) forecasting results are not interpretable for humans. To address these two problems, we propose a novel Knowledge-Driven Temporal Convolutional Network (KDTCN) for stock trend prediction and explanation. Firstly, we extract structured events from financial news, and utilize external knowledge from knowledge graph to obtain event embeddings. Then, we combine event embeddings and price values together to forecast stock trend. We evaluate the prediction accuracy to show how knowledge-driven events work on abrupt changes. We also visualize the effect of events and linkage among events based on knowledge graph, to explain why knowledge-driven events are common sources of abrupt changes. Experiments demonstrate that KDTCN can (i) react to abrupt changes much faster and outperform state-of-the-art methods on stock datasets, as well as (ii) facilitate the explanation of prediction particularly with abrupt changes.

CCS CONCEPTS

• **Information systems** → **Data streams**; **Temporal data**; • **Mathematics of computing** → *Time series analysis*; • **Applied computing** → *Economics*.

KEYWORDS

Knowledge-driven, stock trend prediction, event extraction, structured, unstructured, predictive analytics, explanation

ACM Reference Format:
Shumin Deng, Ningyu Zhang, Wen Zhang, Jiaoyan Chen, Jeff Z. Pan, and Huajun Chen. 2019. Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolutional Network. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3317701

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317701

1 INTRODUCTION

Stock trend prediction has been widely studied due to its scientific and economic merits, and recent efforts mostly focus on exploring potential of deep neural network models. Although such methods [22, 26, 44] perform well in many tasks, they show weakness in tackling the problem of capturing unexpected abrupt changes, and falling short of giving explanations for prediction results.

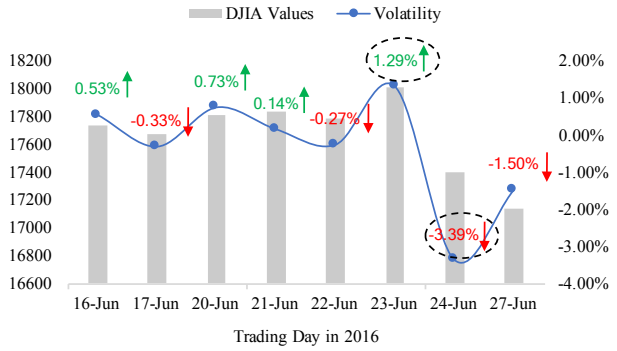


Figure 1: Example DJIA values with abrupt changes

Stock trend prediction with abrupt changes. In stock trend prediction, abrupt changes mean that stock prices fluctuate sharply in an extremely short time interval [7, 16, 19, 28, 43]. For example, as shown in Figure 1, the DJIA (Dow Jones Industrial Average) index increased by 1.29% on 23rd June, 2016, while fell by 3.39% sharply on the next day. For the sake of adapting to abrupt trend changes, [26] have proposed a TreNet to learn both local and global numerical features from stock prices. However, it can be insufficient to merely utilize price data. [17] has shown that stock prices can only reflect all known information, and price movements are in response to news or events. Such as in Figure 1, between 23rd June and 24th June, an unexpected event of *British referendum* happened. In order to encode events for stock trend prediction, [11] have demonstrated

the validity of deep learning methods for event-driven stock market prediction, through event-embedding-based news representations. Despite news events help people capture abrupt changes of stock trend swiftly, they are often disordered and sparse. To address this problem, we import exogenous knowledge to represent events. Knowledge, coming from knowledge graphs (KGs), have two major advantages: (i) enriched semantic information in knowledge can help to establish association among discrete events, and (ii) knowledge in KG is structured and facile to be parameterized.

Deep prediction models lack explanations. Machine learning (ML) explanation, such as interpreting prediction models or justifying prediction results, can significantly increase decision makers' confidence on prediction and boost its application [4, 33]. Even if deep prediction models successfully detect abrupt changes in the stock market, it is hard for them to make people without ML expertise understand why these changes happen. For example, in event-driven stock trend prediction, people may be concerned with which events have greatly influenced stock fluctuation, and how these events take effects. To tackle this problem, we encode interpretable knowledge in deep prediction models, making prediction explainable.

Above all, in this paper, we propose a novel Knowledge-Driven Temporal Convolutional Network (KDTCN), incorporating background knowledge, news events and price data into deep prediction models, to tackle the problem of stock trend prediction and explanation with abrupt changes. We choose Temporal Convolution Network (TCN) [3] because it outperforms canonical RNNs such as LSTMs across a diverse range of tasks and datasets, while demonstrating longer effective memory. The intuition is that *many events are responsible for abrupt changes in the stock market, and the correlation analysis between events and changes offer explanations.*

Specifically, to address the problem of prediction with abrupt changes, we extract events from financial news and structurize them into event tuples, *e.g.*, "*Britain exiting from EU*" is represented as $(Britain, exiting\ from, EU)$. Then entities and relations in event tuples are linked to KGs, such as Freebase [5] and Wikidata [39]. Secondly, we vectorize structured knowledge, textual news, as well as price values respectively, and then concatenate them together. Finally we feed these embeddings into a TCN-based model. Experiments demonstrate that KDTCN can react to abrupt changes in the stock market more swiftly than state-of-the-art methods. Furthermore, based on prediction results with abrupt changes, we address the problem of making explanations. We visualize the effect of events, and also present the linkage among events with the use of KG. By doing so, we make explanations of (i) how knowledge-driven events influence the stock market fluctuation in different levels, and (ii) how knowledge helps to associate events with abrupt changes in stock trend prediction.

To the best of our knowledge, KDTCN is the first one to utilize Temporal Convolution Network to make stock trend prediction, integrating structured knowledge graph, textual news as well as time-series price values. Additionally, KDTCN is capable of explaining prediction results particularly with abrupt changes. In this paper, the next section reviews related work on stock price prediction and ML explanation. Section 3 presents the architecture of KDTCN and specifically introduces the model. Section 4 presents the experiments and evaluation. Section 5 makes a conclusion of the paper and discusses the future work.

2 RELATED WORK

2.1 Deep Models for Stock Prediction

Traditional models of stock prediction are mostly based on sequence modeling with sequence data input. Traditional Recurrent Neural Networks (RNNs) [14, 35, 42] are powerful in discovering the dependency of sequence data, however suffer from vanishing gradients and thus have difficulty in capturing long-term dependencies. Long Short-Term Memory (LSTM) [21] overcomes this limitation. There have already existed some RNN-based stock prediction. [26] have proposed TreNet, a novel end-to-end hybrid neural network, to learn local and global contextual features for predicting stock trend. A dual-stage attention based RNN (DA-RNN) [31], is able to capture long-term temporal dependencies appropriately with attention mechanism. These ML methods with time-series values input, are adapted to the relatively stable stock market, but they have difficulty in reacting to abrupt changes of the stock market swiftly.

Except for value-based models [13, 46], there are also some methods using texts. [22] have proposed hybrid attention networks to predict stock trend based on the sequence of recent news. [10, 11] have extracted events from news and demonstrated that deep learning is useful for event-driven stock movement prediction. Although these models convert unstructured text to structural events, they only utilize texts, making performance limited. [1, 44] use both numerical and textual data. [1] have modeled temporal effects of past events on opening prices with LSTM. [44] have presented a deep generative model to predict stock movement from tweets and historical stock prices. However, these two models are unable to represent chaotic social text effectively, as they only utilize bag-of-words or word embeddings without capturing structured relations.

2.2 Knowledge-driven Models and Explanation

Despite substantial efforts have been made for stock prediction, most of them only learn features from numerical and textual data, while ignore background knowledge. [9] have demonstrated that incorporating knowledge can help capture inconsistent evolution of stream data, thereby make more accurate prediction. [12] have proposed to incorporate KG into the learning process of event embeddings, which can encode valuable background knowledge. Besides, knowledge-driven ML models also show their strengths in other broad domains, such as recommender systems. [23] have proposed a novel knowledge enhanced sequential recommender, integrating RNN-based networks with Key-Value Memory Network (KV-MN). [41] have proposed a deep knowledge-aware network (DKN), incorporating KG to news recommendation.

In these knowledge-driven models, eXplainable AI (XAI) is very important. In ML literature, work on explanation often focuses on visualizations of prediction. Beyond that, research focuses on two broad approaches to explanation [4]. The first is prediction justification, where a (usually non-interpretable) model [8, 18, 38] and prediction are given, and a justification for the prediction must be produced. The second is interpretable models, aiming to devise models that are intrinsically interpretable and can be explained by reasoning. In this paper, we focus on the first one. There are many producing justifications, and they focus on interpreting predictions of specific complex models, often by proposing to isolate contributions of individual features. [33] have proposed to explain models

by presenting representative individual predictions and their explanations in a non-redundant way. They have explained predictions of each classifier by learning an interpretable model locally.

3 METHODOLOGY

In this section, we first present the overview of the proposed knowledge-driven temporal convolutional network (KDTCN) for stock trend forecasting and explanation. Then we introduce each model component in detail.

3.1 Model Overview

The overview of KDTCN architecture is shown in Figure 2. Original model inputs are price values $\mathcal{X}$, news corpus $\mathcal{N}$, and knowledge graph $\mathcal{G}$. The price values are normalized and mapped into the price vector, denoted by

$$\mathcal{P} = \{p_0, p_1, \cdots, p_{T-1}\}, \tag{1}$$

where each vector p_t represents a real-time price vector on a stock trading day t, and T is the time span.

In news corpus, pieces of news are represented as event sets $\mathcal{E}$, and are structurized by open domain information extraction (Open IE [15]) through leveraging linguistic structure. Each event is structurized into an event tuple $e = (s, p, o)$, where p is the action or predicate, s is the actor or subject and o is the object on which the action is performed. Then, each item in event tuples is linked to KG. Note that event items in this paper refer to the s, p and o in the event tuple (s, p, o), and they also correspond to entities and relations in KG. We obtain event embeddings $\mathcal{V}$ by training both event tuples and KG triples. Finally, event embeddings, combined with price vectors are input into a TCN-based model [3] for stock trend prediction and explanation particularly with abrupt changes.

Our purpose is to forecast the movement of target stock index trend y with abrupt changes. We predict the binary movement, in which 1 denotes `rise` and 0 denotes `drop`, defined by

$$y = \begin{cases} 1 & if \;\; x_t > x_{t-1} \\ 0 & otherwise \end{cases} \tag{2}$$

where x_t denotes the stock price value on the stock trading day t.

3.2 Event Embedding

The goal of event embedding is to learn low-dimension dense vector representations for event tuples $e = (s, p, o)$. We first extract structured event tuples from financial news, and then link them to KG. An event tuple embedding is calculated by multi-channel concatenation of KG embeddings and word vectors for each item.

3.2.1 Event Extraction and Structuralization. We convert unstructured news texts into structured event tuples by Open IE [15], the goal of which is to read a sentence and extract tuples with a relation phrase and arguments that are related by that relation phrase. Originally, Open IE extracts binary tuples [29], *i.e.*, two arguments connected by one relation phrase. *E.g.*, "*Britain exiting from the EU*", with the subject-predicate-object structure, is structured into an event tuple $(s = Britain, p = exiting\ from, o = EU)$.

After event extraction based on Open IE, there can be lots of redundancy in generated event tuples. Thus we have also removed useless words (*e.g.*, adjectives and adverbs) in fundamental sentence structures, to ensure that event tuples $\mathcal{E}$ are concise enough.

3.2.2 Entity Linking and Extension. After getting concise event tuples, we construct a sub-graph from KG by utilizing the technique of entity linking [36], in order to disambiguate named entities in texts by associating them with predefined entities in KG. Note that the subject s, predicate p and object o in an event tuple may not always have linkage in KG. Besides, information in a single event tuple may be sparse and lack diversity. Thus, we enrich the sub-graph by importing immediate neighbors of linked entities within one hop in KG. To formulate these concepts, we propose $linking(e)$ and $linking(r)$ to define the entity and relation in an event tuple linked to KG, as well as $context(e)$ to define the immediate neighbors of linked entities in KG, denoted by

$$linking(e) = \{e_i \mid e_i = s \vee o,\ (s, p, o) \in \mathcal{E}_t \ \wedge \ e_i \in \mathcal{G}\} \tag{3}$$

$$linking(r) = \{r_i \mid r_i = p \ \wedge \ (s, p, o) \in \mathcal{E}_t \ \wedge \ r_i \in \mathcal{G}\} \tag{4}$$

$$context(e) = \{e_i \mid (e_i, r, e) \in \mathcal{G} \ \vee \ (e, r, e_i) \in \mathcal{G}, e \in linking(e)\} \tag{5}$$

3.2.3 Knowledge-driven Multi-channel Concatenation. We choose TransE [6] as the KG embedding method in this paper, as it could generally preserve structural information in KG with great robustness. As mentioned before, not all entities and relations in event tuples can be linked to KGs. In these situations, zero padding [40] is chosen. If $s, o \notin linking(e)$, $\mathcal{V}_{e_l^s} = \vec{\mathbf{0}}$ and $\mathcal{V}_{e_l^o} = \vec{\mathbf{0}}$. If $p \notin linking(r)$, $\mathcal{V}_{r_l^p} = \vec{\mathbf{0}}$. Besides, the context of each linked entity tend to be more than one entity and relation, so the context embedding is calculated by averaging. We then parameterize event representations in different channels, denoted by $\mathcal{V}_l$ in the channel of KG linking, $\mathcal{V}_c$ in the channel of KG context, and $\mathcal{V}_w$ in the channel of words.

$$\mathcal{V}_l = [\mathcal{V}_{e_l^s}, \mathcal{V}_{r_l^p}, \mathcal{V}_{e_l^o}] \tag{6}$$

$$\mathcal{V}_c = [\mathcal{V}_{e_c^s}, \mathcal{V}_{r_c^p}, \mathcal{V}_{e_c^o}] \tag{7}$$

$$\mathcal{V}_w = [\mathcal{V}_{e_w^s}, \mathcal{V}_{r_w^p}, \mathcal{V}_{e_w^o}] \tag{8}$$

where $e_l^s, e_l^o \in linking(e)$, and $r_l^p \in linking(r)$; $e_c^s, e_c^o \in context(e)$, and $r_c^p \in \mathcal{G}$; $\mathcal{V}_{e_w^s}$, $\mathcal{V}_{r_w^p}$, and $\mathcal{V}_{e_w^o}$ are the word vectors of s, p, and o respectively; $\mathcal{V}_*$ represents the embedding of $*$.

Then we concatenate $\mathcal{V}_l$, $\mathcal{V}_c$ and $\mathcal{V}_w$ in multiple channels to get the final event embedding, denoted by

$$\mathcal{V} = F_e(\mathcal{E}, \mathcal{G}) = [\mathcal{V}_l\ \mathcal{V}_c\ \mathcal{V}_w] \tag{9}$$

3.3 Temporal Convolutional Network

We refer to the presented TCN architecture proposed by [25, 32], and note that the basic TCN model we adopt in this paper is based on a generic architecture described by [3].

TCN [3, 25, 32] uses a 1-D fully-convolutional network (FCN) architecture [27], where each hidden layer has the same length as the input layer, and zero padding [40] is added to keep subsequent layers the same length as previous ones. In this way, the network can produce an output of the same length as the input. Besides, TCN

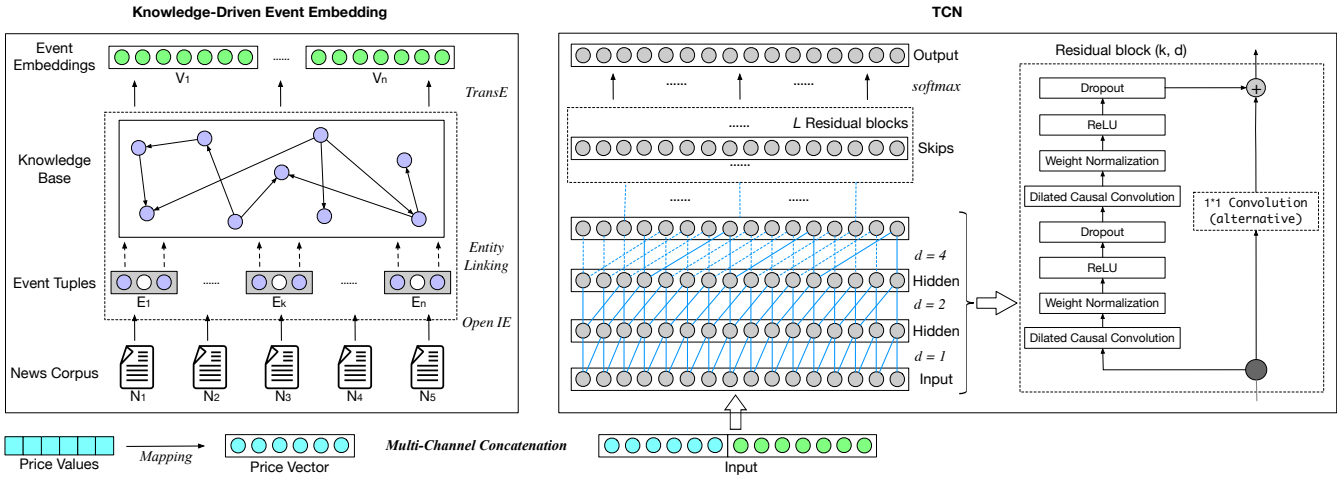


Figure 2: Illustration of the KDTCN framework. (Best viewed in color.)

uses causal convolutions, where an output at time t is convolved only with elements from time t and earlier in the previous layer, so that there can be no leakage from the future to the past.

Simply put, TCN = 1-D FCN + causal convolutions.

Then, in the following, we describe how techniques from present convolutional architectures are integrated into a TCN, considering both deep networks and long-range dependence.

3.3.1 Dilated Convolutions. Formally, for a 1-D sequence input $X \in \mathbb{R}^n$ and a filter $\mathcal{F}$, the dilated convolution operation on j_{th} element in the sequence X is defined as

$$O(X_j) = (X *_d \mathcal{F}) = \sum_{i=0}^{k-1} \mathcal{F}(i) * X_{X_j - i \cdot d}, \tag{10}$$

where d is the dilation factor, k is the filter size, and the subscript $X_j - i \cdot d$ denotes the direction of the past.

In fact, dilation can be regarded as importing a fixed step between every two adjacent filter taps. Each layer consists a set of dilated convolutions with rate parameter d, a non-linear activation $f(\cdot)$, and a residual connection that combines the layer's input and the convolution signal. d increases consecutive layers within a block, calculated by $d_l = 2^l$. Convolutions are only applied over two timestamps, t and $t - d$. Specifically, filters can be parameterized by weight matrices $W = [W_0, W_1]$ and the bias vector b, where $W_i \in \mathbb{R}^{F_w \times F_w}$, $b \in \mathbb{R}^{F_w}$, and F_w denotes the number of filters.

$\tilde{Z}_t^{(j,l)}$ and $Z_t^{(j,l)}$ are results after dilated convolution and adding the residual connection at timestamp t respectively, denoted by

$$\tilde{Z}_t^{(j,l)} = f(W_0 \tilde{Z}_{t-d}^{(j,l-1)} + W_1 \tilde{Z}_t^{(j,l-1)}), \tag{11}$$

and

$$Z_t^{(j,l)} = Z_t^{(j,l-1)} + V \tilde{Z}_t^{(j,l)} + e, \tag{12}$$

where $V \in \mathbb{R}^{F_w \times F_w}$ denotes the weight matrix and $e \in \mathbb{R}^{F_w}$ denotes the bias vector for the residual block.

3.3.2 Residual Connections. [20] has presented that a residual learning framework can ease network training, indicating residual blocks benefit very deep networks. Referring to [3, 32], we define a series of residual blocks, each of which contains a series of L convolutional layers. Activations in the l_{th} layer and j_{th} block are $Z^{(l,j)} \in \mathbb{R}^{F_w \times T}$, where T is the time span, and F_w is same at each layer. The calculation of $Z^{(l,j)}$ is defined in Equation (12).

Within a residual block, the TCN has two layers of dilated causal convolution and non-linearity, and we use ReLU [30]. For normalization, we apply weight normalization to convolutional filters. In addition, a spatial dropout [37] was added after each dilated convolution for regularization: at each training step, a whole channel is zeroed out.

However, whereas in standard ResNet [20] the input is added directly to the output of the residual function, in TCN (and ConvNets in general) the input and output could have different widths. To account for discrepant input-output widths, we use an additional 1×1 convolution to ensure element-wise addition $\oplus$ receiving tensors of the same shape.

3.4 Interpretation of Event Effects

We adopt a well-explored concept, *effect*, in the prediction, meaning its contribution towards or against the predicted class [34, 45]. In this paper, stock trend prediction can be regarded as the problem of binary classification. To make it simple, the discriminant function for data instance (event tuple e) in the binary-class classifier is

$$f(y, e) = \sum_i \theta_i e_i \tag{13}$$

where e_i denotes the instance value of events, and the weight coefficients θ_i is learned from training data for each class y.

Then the classifier predicts the class of the event instance as the one that maximizes the predictor function, through a monotonic non-linear distortion function φ:

$$\hat{y} = \arg\max_y \varphi(f(y, e)) \tag{14}$$

So the effect of the i_{th} event towards or against predicting class y for a data instance can be denoted as

$$F_{ef_i} = \theta_i e_i \tag{15}$$

4 EXPERIMENTS

The experiments mainly consist of two parts: (i) prediction evaluation and (ii) case-based explanation for prediction.

4.1 Datasets & Baselines & Settings

Datasets. Datasets in this paper are listed below.

- **Time-series Price Data $\mathcal{X}$**: The price dataset contains daily value records of DJIA index, and also has timespan from 08/08/2008 to 01/01/2016. Stock price data are downloaded directly from Yahoo Finance[1]. Note that they are cleaned for bank holidays, and aligned with financial news by time.
- **Textual News Data $\mathcal{N}$**: The news dataset is composed of historical news headlines from Reddit WorldNews Channel[2]. For each stock trading day, the top 25 news headlines are selected based on Reddit users' votes. The timespan is from 08/08/2008 to 01/01/2016.
- **Structured Knowledge Data $\mathcal{G}$**: The structured data come from two commonly used open knowledge graphs for research, including the Freebase [5] and the Wikidata [39]. We construct a sub-graph based on them, which contains 64958 entities and 716 relations totally.

Baselines. We consider commonly-used baseline model variations as shown in Table 1.

Model	Input	
	Raw Data	Processed Training Data
ARIMA [2]	$\mathcal{X}$	$\mathcal{X}$
LSTM	$\mathcal{X}$	$\mathcal{X}$
CNN	$\mathcal{X}$	$\mathcal{X}$
TCN	$\mathcal{X}$	$\mathcal{X}$
WB-TCN	$\mathcal{N}$	*word embedding*
EB-TCN	$\mathcal{N}$	*event embedding*$^{(a)}$
PVWB-TCN	$\mathcal{X}$ + $\mathcal{N}$	*price vector + word embedding*
PVEB-TCN	$\mathcal{X}$ + $\mathcal{N}$	*price vector + event embedding*$^{(a)}$
KDEB-TCN	$\mathcal{N}$ + $\mathcal{G}$	*event embedding*$^{(b)}$
KDTCN	$\mathcal{X}$ + $\mathcal{N}$ + $\mathcal{G}$	*price vector + event embedding*$^{(b)}$

Table 1: Baseline models with different inputs. In the first column, prefix WB means word embeddings, EB means event embeddings, PV means the price vector, and KD means knowledge-driven. Note that *event embedding*$^{(a)}$ and *event embedding*$^{(b)}$ denote event embedding without and with KG respectively.

Settings. With regard to settings of training process, stochastic gradient descent (SGD) [24] optimizer is used, with k kernels and L levels residual blocks. The best performance is gotten when $k = 2$ and $L = 10$. The dimension of hidden units is 100, and the dimension of word embedding, entity embedding as well as relation embedding are all set to 50. In KDTCN, a dropout rate of 0.5 is used to avoid over-fitting, and the learning rate is 1×10^{-5}. We split the raw value dataset into training set and testing set with ratio 0.8 and 0.2. We evaluate prediction performance with two different evaluation metrics, including (i) *Accuracy* and (ii) *F1 score* .

[1]https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI
[2]https://www.reddit.com/r/worldnews/?hl

4.2 Prediction Evaluation

Performance of KDTCN is shown in three progressive aspects: (i) evaluation of basic TCN architecture, (ii) influence of different model inputs with TCN, and (iii) TCN-based model performance for abrupt changes.

4.2.1 Basic Evaluation for TCN. In order to demonstrate that generic TCN architecture can outperform some traditional prediction models, we make comparisons with them, shown in Table 2. Note that all experiments reported in this part are only input with price values.

Model	Accuracy	F1 Score
ARIMA	52.40% ± 0.25%	0.3844 ± 0.0052
LSTM	56.17% ± 0.26%	**0.5348 ± 0.0039**
CNN	55.41% ± 0.25%	0.5280 ± 0.0021
TCN	**60.96% ± 0.26%**	0.5317 ± 0.0036

Table 2: Stock trend prediction results over the DJIA index dataset with different basic prediction models.

In Table 2, we observe that TCN greatly outperforms baseline models on the stock trend prediction task. TCN achieves much better performance than either traditional ML models (such as ARIMA), or deep neural networks (such as LSTM and CNN), indicating that TCN has more obvious advantages in sequence modeling and classification problems. Therefore, we choose TCN as our basic prediction model in this paper.

4.2.2 Different Model Inputs with TCN. For the sake of validating effectiveness of integrating knowledge graph, financial news corpus, and price values in stock trend prediction, we compare prediction performance of models with different inputs, shown in Table 3.

Model	Accuracy	F1 Score
TCN	60.96% ± 0.26%	0.5317 ± 0.0036
WB-TCN	62.96% ± 0.25%	0.5934 ± 0.0038
EB-TCN	64.74% ± 0.26%	0.6698 ± 0.0031
PVWB-TCN	66.49% ± 0.24%	0.5845 ± 0.0043
PVEB-TCN	67.51% ± 0.25%	0.6979 ± 0.0031
KDEB-TCN	68.51% ± 0.25%	0.6959 ± 0.0032
KDTCN	**69.77% ± 0.25%**	**0.7073 ± 0.0030**

Table 3: Stock trend prediction results over the overall DJIA index dataset with different inputs on TCN-based models.

As seen, WB-TCN and EB-TCN both get better performance than TCN, indicating textual information helps to improve forecasting. Analogously, compared PVWB-TCN and PVEB-TCN with WB-TCN

and EB-TCN, the comparison results reflect that price values are also useful in stock trend prediction. KDEB-TCN outperforms other baselines, which accounts for structured knowledge greatly working on stock trend prediction. Moreover, KDTCN gets both the highest *accuracy* and *F1 scores*, and such a result demonstrates the validity of model input integration with structured knowledge, financial news, and price values.

4.2.3 Model Performance for Abrupt Changes. We intend to verify if knowledge-driven models can effectively capture abrupt changes of stock trend prediction in this part. We first get time intervals of abrupt changes by figuring out the difference of stock fluctuation degree $\mathcal{D}_{fluctuation}$ between two adjacent stock trading days [22], calculated by

$$\mathcal{D}_{fluctuation} = \frac{x_t - x_{t-1}}{x_{t-1}}, \tag{16}$$

where x_t and x_{t-1} denotes the stock price value on the stock trading day t and $t-1$ respectively. Then the difference of fluctuation degree C is defined by:

$$C_i = \mathcal{D}_{fluctuation_i} - \mathcal{D}_{fluctuation_{i-1}}. \tag{17}$$

Intuitively, the larger $|C_i|$ is, the more likely for the i_{th} day locates in the time interval of abrupt changes. If $|C_i|$ exceeds a certain threshold, it can be considered that the stock price abruptly changes at the i_{th} day. In order to identify a proper range of thresholds, we show the performance based on data distribution of $|C|$ in Figure 3.

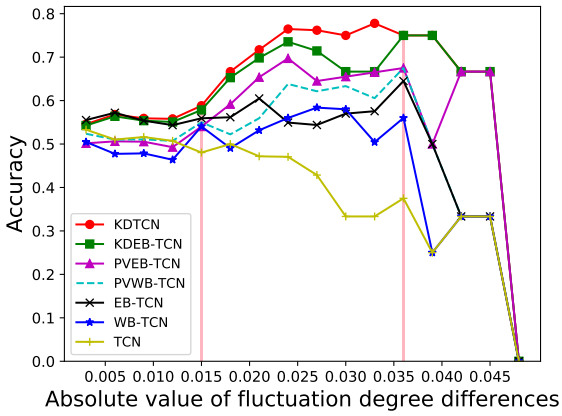


Figure 3: Performance on the data distribution of $|C|$.

Seen from Figure 3, overall, higher accuracy is achieved when $|C|$ are in the interval of 0.015 to 0.036, except for TCN marked in olive. TCN achieves the much worse performance, demonstrating that the numerical-data-based stock prediction model may perform poorly on prices with abrupt changes. Besides, KDTCN achieves more stable and better performance than other baselines, showing advantages of knowledge-driven models with integration of knowledge, texts and values. Then we calculate the average *Accuracy* and *F1 score* of KDTCN and baselines on datasets with abrupt changes, where $|C|$ are in the interval of 0.015 to 0.036, presented in Table 4.

Model	Accuracy	F1 Score
ARIMA	30.18% ± 0.25%	∞
LSTM	36.27% ± 0.25%	∞
CNN	38.33% ± 0.25%	∞
TCN	43.11% ± 0.25%	∞
WB-TCN	54.16% ± 0.24%	0.5078 ± 0.0038
EB-TCN	56.64% ± 0.26%	0.5882 ± 0.0038
PVWB-TCN	59.56% ± 0.24%	0.5556 ± 0.0031
PVEB-TCN	63.54% ± 0.25%	0.6316 ± 0.0033
KDEB-TCN	67.32% ± 0.26%	0.6667 ± 0.0032
KDTCN	**71.80% ± 0.25%**	**0.7388 ± 0.0030**

Table 4: Stock trend prediction results over the local DJIA index dataset of abrupt changes, with different model inputs.

We observe that models with knowledge-driven event embedding input, such as KDEB-TCN and KDTCN, can greatly outperform numerical-data-based and textual-data-based models. These comparison results indicate that knowledge-driven events have significant impacts on the stock trend prediction with abrupt changes, and knowledge-driven models have advantages in reacting to abrupt changes in the stock market swiftly. Furthermore, KDTCN achieves better performance than KDEB-TCN, demonstrating that integrating price data to knowledge-driven models can also benefit stock trend prediction with abrupt changes.

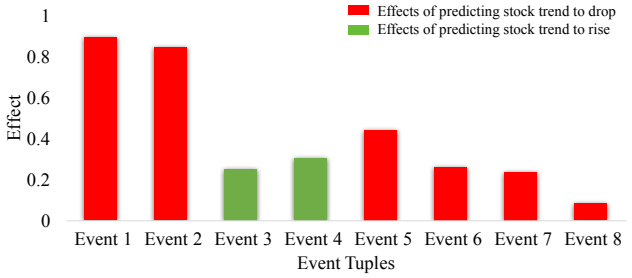
4.3 Explanation for Prediction

Explanation in this paper belongs to human-centric justification of ML predictions [4]. We explain why knowledge-driven events are common sources of abrupt changes to human without ML expertise. The explanations are accomplished in two aspects: (i) visualizing effects of knowledge-driven events on prediction results with abrupt changes, and (ii) retrieving background facts of knowledge-driven events by linking the events to external KG.

4.3.1 Effect Visualization of Events. We calculate different event effects on stock trend prediction with Equation (14) and (15), then we visualize the results in Figure 4.

The prediction result in Figure 4 is that trend of DJIA index will `drop`. Note that the bars of the same colour have the same event effect, the height of bars reflects the degree of effects, and the event popularity[3] declines from left to right. Intuitively, events with higher popularity should have greater effects on stock trend prediction with abrupt changes, but not always. As seen, events of *Brexit* and *EU Referendum* both play an important role in forecasting downward trends. Nearly all other events with negative effect are related to these two events, *e.g.*, (*British Pound, drops, nearly 5%*) and (*Northern Ireland, calls for poll on United Ireland*). Although there are also some events have positive effects of predicting stock trend to rise, and have high popularity, *i.e.*, (*Rich, Getting, Richer*), the total effect is negative. Therefore, abrupt changes of the stock index fluctuation can be viewed as the combined result of effects and popularity of events.

[3]The event popularity is obtained through the ranking of top 25 news headlines which are selected based on Reddit users' votes.



Event 1	[Britain, existing from, EU]
Event 2	[United Kingdom, votes to leave, European Union]
Event 3	[Rich, Getting, Richer]
Event 4	[Buenos Aires zoos 2,500 animals, moved to, nature reserves]
Event 5	[British Pound, drops, nearly 5%]
Event 6	[J. K. Rowlin, leads the charge for, Scottish independence]
Event 7	[Northern Ireland, calls for poll on United Ireland]
Event 8	[London banking jobs, start to, cut]

Figure 4: Examples of event effect on stock trend prediction.

4.3.2 Visualization of Event Tuples Linked to KG. We present the sample KG triples linked to event tuples in Figure 5.

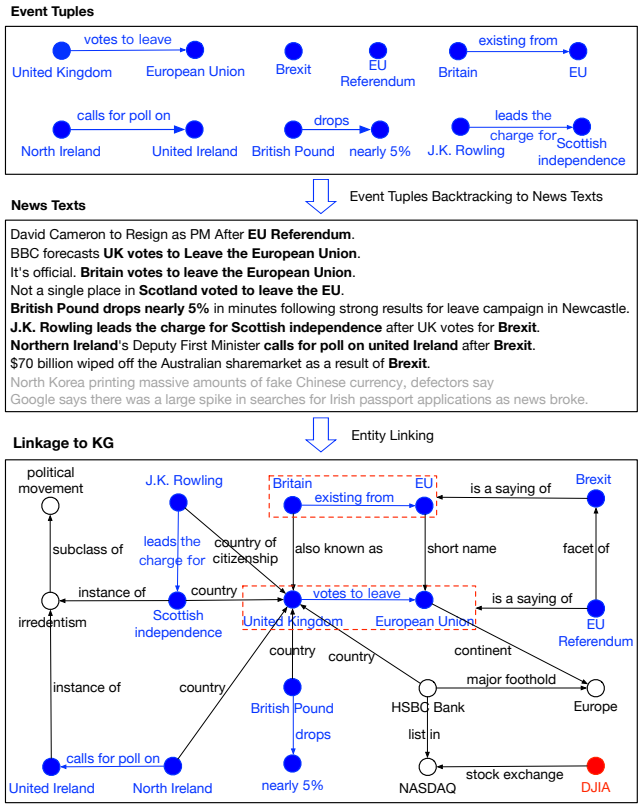


Figure 5: Illustration of triples in KG linked to events.

First, we search the event tuples with great effects or high popularity in stock trend movements. Then, we backtrack to the news texts containing these events. Finally, we retrieve associated KG triples linked to event tuples by entity linking. In Figure 5, each event tuple is marked in blue, and entities in it are linked to KG. We also mark *DJIA* in red. As seen, we illustrate that knowledge can help to associate events with abrupt changes in stock trend prediction together. These listed event tuples, such as (*Britain, existing from, EU*), (*United Kingdom, votes to leave, European Union*), (*British Pound, drops, nearly 5%*), (*J. K. Rowlin, leads the charge for, Scottish independence*), and (*Northern Ireland, calls for poll on United Ireland*), are not strongly relevant literally. However, with the linkage to KG, they can establish association with each other, and strongly related to events of *Brexit* and *EU Referendum*. Besides, these knowledge-enhanced events can also have connections with *DJIA*, thus originally sparse events can be closely linked and have combined effects on *DJIA* index. Thus, examples in Figure 5 explain how knowledge-driven events work on stock movements, and why knowledge-driven models are valid. Moreover, by incorporating explanations of event effects, we justify that knowledge-driven events are common sources of abrupt changes.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel knowledge-driven temporal convolutional network (KDTCN) to tackle the problem of stock trend prediction and explanation with abrupt changes. We extract structured event tuples from financial news, and utilize background knowledge from KG to associate discrete event tuples with each other. Through training both event tuples and KG triples, we get knowledge-driven event embeddings. Furthermore, we integrate price vectors and event embeddings as prediction model inputs by multi-channel concatenation. We utilize TCN to predict stock trend, and also explain prediction results based on knowledge. The experiments on stock datasets demonstrate that integrating structured knowledge to TCN can (i) greatly outperform present deep models when forecasting stock trend with abrupt changes, and (ii) make explanation on prediction results with abrupt changes. Through the event effect visualization and knowledge-enhanced event tuple visualization, we explain how knowledge influences greatly on stock trend with abrupt changes.

Based on research in this paper, we have identified several potential directions of this work, mainly includes a more general evaluation of different event effects on stock trend, a study on long-range dependency of events, and a more specific experiment on prediction explanations. At a stock trading day, there are various events influencing stock movements. We will figure out different effects of them and categorize these events based on effects, for example, which type of events affects stock trend to a great extent. Beside, the effect of an event may change as time goes by, thus to capture the dynamic effects of events is valuable. Furthermore, in this paper, we only give a case-based experiment on how knowledge help to establish association among sparse events. In the future, we will propose some quantitative indices to evaluate the effectiveness of knowledge, and give more specific explanations.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge that this work is funded by NSFC 61473260/91846204, national key research program YS2018YFB140004 as well as Natural Science Foundation of Zhejiang Province of China (LQ19F030001), and supported by Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

REFERENCES

[1] Ryo Akita, Akira Yoshihara, Takashi Matsubara, and Kuniaki Uehara. 2016. Deep learning for stock prediction using numerical and textual information. In *Computer and Information Science (ICIS), 2016*. IEEE, 1–6.

[2] Dimitrios Asteriou and Stephen G. Hall. 2011. *ARIMA Models and the Box-Jenkins Methodology*.

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. (2018).

[4] Or Biran and Kathleen R McKeown. 2017. Human-Centric Justification of Machine Learning Predictions.. In *IJCAI*. 1461–1467.

[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 1247–1250.

[6] Antoine Bordes, Nicolas Usunier, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *International Conference on Neural Information Processing Systems*. 2787–2795.

[7] Marco Antonio Leonel Caetano and Takashi Yoneyama. 2007. Characterizing abrupt changes in the stock prices using a wavelet decomposition method. *Physica A: Statistical Mechanics and its Applications* 383, 2 (2007), 519–526.

[8] Emilio Carrizosa, Belen Martin, and Dolores Morales. 2006. *A column generation approach for support vector machines*. Technical Report. Technical report.

[9] Freddy LeCue Chen, Jiaoyan and, Jeff Z. Pan, and Huajun Chen. 2017. Learning from Ontology Streams with Semantic Concept Drift. In *IJCAI*. 957–963.

[10] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of EMNLP 2014*. 1415–1425.

[11] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction.. In *Ijcai*. 2327–2333.

[12] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In *Proceedings of COLING 2016*. 2133–2142.

[13] Aaron Elliot, Cheng Hua Hsu, and Jennifer Slodoba. 2017. Time Series Prediction: Predicting Stock Price. *arXiv preprint arXiv:1710.05751* (2017).

[14] Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7, 2-3 (1991), 195–225.

[15] Oren Etzioni, Michael Cafarella, and Michele Banko. 2014. OPEN INFORMATION EXTRACTION.

[16] Eugene F Fama. 1963. Mandelbrot and the stable Paretian hypothesis. *The journal of business* 36, 4 (1963), 420–429.

[17] Eugene F Fama. 1965. The behavior of stock-market prices. *The journal of Business* 38, 1 (1965), 34–105.

[18] Raphael Féraud and Fabrice Clérot. 2002. A methodology to explain neural network classification. *Neural Networks* 15, 2 (2002), 237–246.

[19] James D Hamilton and Gang Lin. 1996. Stock market volatility and the business cycle. *Journal of Applied Econometrics* 11, 5 (1996), 573–593.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015), 770–778.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[22] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 261–269.

[23] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 505–514.

[24] Nikhil Ketkar. 2014. Stochastic Gradient Descent. *Optimization* (2014).

[25] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*. Springer, 47–54.

[26] Tao Lin, Tian Guo, and Karl Aberer. 2017. Hybrid Neural Networks for Learning the Trend in Time Series. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*. 2273–2279.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[28] J. I. Maskawa. 2016. Collective Behavior of Market Participants during Abrupt Stock Price Changes. *Plos One* 11, 8 (2016), e0160152.

[29] Mausam Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 4074–4077.

[30] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on International Conference on Machine Learning*. 807–814.

[31] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. (2017), 2627–2633.

[32] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *IEEE International Conference on Computer Vision (ICCV)*.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[34] Marko Robnik-Šikonja and Igor Kononenko. 2008. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering* 20, 5 (2008), 589–600.

[35] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536.

[36] Avirup Sil and Alexander Yates. 2013. Re-ranking for joint named-entity recognition and linking. 18, 5 (2013), 2369–2374.

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[38] Henri Jacques Suermondt. 1992. Explanation in Bayesian belief networks. (1992).

[39] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[40] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. 1990. Phoneme recognition using time-delay neural networks. In *Readings in speech recognition*. Elsevier, 393–404.

[41] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep Knowledge-Aware Network for News Recommendation. *arXiv preprint arXiv:1801.08284* (2018).

[42] P.J. Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 10 (1990), 1550–1560.

[43] Eugene N White. 1990. The stock market boom and crash of 1929 revisited. *Journal of Economic perspectives* 4, 2 (1990), 67–83.

[44] Yumo Xu and Shay B Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1970–1979.

[45] Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang. 2008. Explaining inferences in Bayesian networks. *Applied Intelligence* 29, 3 (2008), 263–278.

[46] Liheng Zhang, Charu Aggarwal, and Guo Jun Qi. 2017. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2141–2149.

WebProtégé: A Cloud-Based Ontology Editor

Matthew Horridge
Stanford University
matthew.horridge@stanford.edu

Rafael S. Gonçalves
Stanford University
rafael.goncalves@stanford.edu

Csongor I. Nyulas
Stanford University
nyulas@stanford.edu

Tania Tudorache
Stanford University
tudorache@stanford.edu

Mark A. Musen
Stanford University
musen@stanford.edu

ABSTRACT

We present WebProtégé, a tool to develop ontologies represented in the Web Ontology Language (OWL). WebProtégé is a cloud-based application that allows users to collaboratively edit OWL ontologies, and it is available for use at https://webprotege.stanford.edu. WebProtégé currently hosts more than 68,000 OWL ontology projects and has over 50,000 user accounts. In this paper, we detail the main new features of the latest version of WebProtégé.

KEYWORDS

OWL; knowledge representation; ontologies; ontology engineering

ACM Reference Format:
Matthew Horridge, Rafael S. Gonçalves, Csongor I. Nyulas, Tania Tudorache, and Mark A. Musen. 2019. WebProtégé: A Cloud-Based Ontology Editor. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3317707

1 INTRODUCTION

The Protégé team at the Stanford Center for Biomedical Informatics Research has been developing software for formally representing knowledge in the form of ontologies since the early 90s [3]. With the growing adoption of logic-based knowledge representation languages that have precisely-defined semantics—such as the World Wide Web Consortium (W3C) standard Web Ontology Language (OWL) [1]—the WebProtégé [2, 5] software has enjoyed increasing traction and demand from its growing user base. In particular, the recent popularity of so-called "knowledge graphs" has brought a new wave of users to the WebProtégé user community, as well as students to our Protégé courses. Ontologies can be used as components of knowledge graphs, to equip them with precise semantics and to enable logical reasoning. We co-developed multiple ontologies that are used in enterprise knowledge graphs, and some of the features in WebProtégé we present here were driven by those experiences. With this paper, we hope to draw feedback from the knowledge graph community so that we can steer our development efforts to better suit the needs of this community.

The WebProtégé cloud-based tool supports simultaneous collaborative editing of OWL ontologies for geographically distributed users. By design, WebProtégé offers a simple user interface streamlined for common ontology engineering patterns and tasks. WebProtégé is available for general purpose use at https://webprotege.stanford.edu, where it hosts over **68,000 OWL ontology projects** and over **50,000 user accounts**. The software is open-source and all source code is available on GitHub.[1] In its latest version, WebProtégé 4.0 builds on the success of previous versions but features a new modern interface, and new features for collaboration, search, and quality assurance.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317707

2 WEBPROTÉGÉ CLOUD ONTOLOGY EDITOR

WebProtégé is essentially a "Google Docs" for editing ontologies. Users access "projects" that are collections of OWL ontologies augmented with a change history, and issues/comments. WebProtégé provides a default user interface that supports lightweight ontology editing, for ontologies that more or less fit within the OWL2EL profile.[2] However, it can also be configured for editing ontologies that require full-blown OWL 2 axioms[3] and class expressions.

2.1 Ontology Editing

Figure 1 shows the main editing interface of WebProtégé, which provides a tabbed look and feel. The classes tab comprises four resizable views: Class Hierarchy, Class Description, Comments and Project Feed. Users can add, remove and customize the layout of views on a tab, and the state of the user interface is automatically persisted for each user in the context of a project.

In contrast to previous versions, WebProtégé 4.0 uses proper URL routing, meaning that entities, such as classes, properties and individuals, and the tabs (or pages) that display them, are bookmarkable links that can be shared with people.

WebProtégé allows ontology edits to be seen by all users as they take place. It is possible to set up a project so that users may only view or comment on a project for cases where there is a core group of editors and a larger group of commenters/reviewers.

Users of WebProtégé can easily link to entities in popular knowledge bases such as **Schema.org**, **Wikidata**, and **DBpedia**. For example, to link out to the entity Dataset from Schema.org, a user types in schema:Dataset when creating a new class and WebProtégé renders the entity to its IRI https://schema.org/Dataset, and similarly for Wikidata and DBpedia entities using prefix names wikidata and dbpedia, respectively.

[1] https://github.com/protegeproject/webprotege
[2] https://www.w3.org/TR/owl2-profiles/#OWL_2_EL
[3] https://www.w3.org/TR/owl2-syntax/

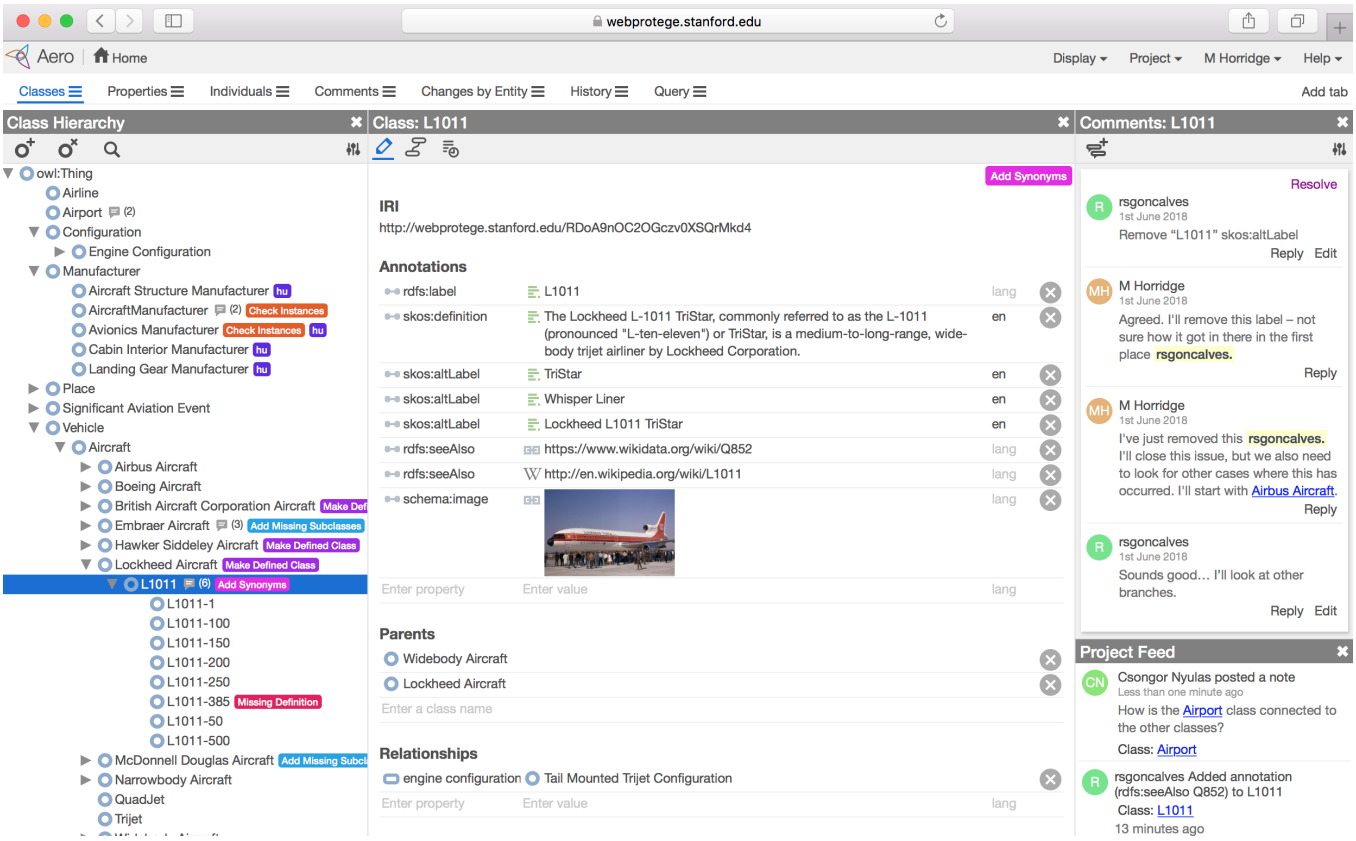


Figure 1: Class editing in WebProtégé. This figure demonstrates the default user interface and shows the use of tags, comments/issues and the project feed.

2.2 Collaborative Features

Besides collaborative ontology editing functionality, WebProtégé also allows users to comment on a project and create issues in the form of *threaded comments*. Comments sit outside of an ontology but pertain to specific entities in an ontology. They are formatted in Markdown syntax and can contain links to any entities in the ontology as well as GitHub style @user mentions.

Discussion threads of comments for a given entity can be seen on the right-hand side of Figure 1 in the Comments view. The number of threads is also displayed next to entity names in the various entity hierarchies (left hand side of Figure 1). Threads can be created and closed as issues are opened and then dealt with by ontology editors. Figure 2 shows a global view of the Comments tab that is available in the default interface.

When comments are posted, project participants are notified via email (left inset in Figure 2). Emails contain the body of the comment and clickable links that take users directly to the comment in a browser window. WebProtégé also provides integration with *Slack*[4] and makes it possible to associate a *Slack* Webhook[5] with a project so that comments are posted to a Slack channel for external

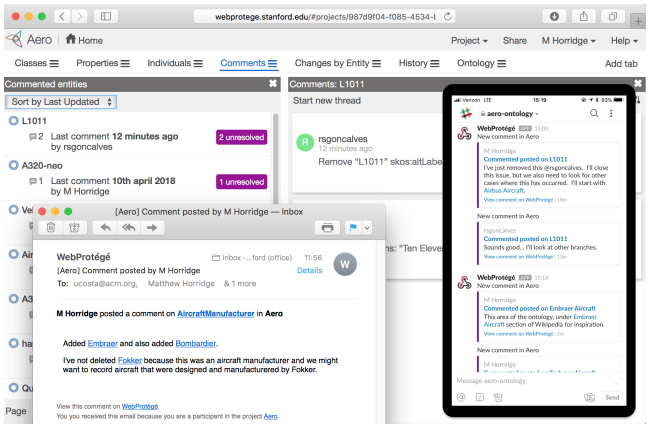


Figure 2: Comments and Issues in WebProtégé. The main part of the figure shows the Comments tab. Comments can be sorted and viewed by entity. The lower left-hand side inset shows a notification email sent out to participants after a comment has been posted. The right-hand side inset shows integration with the chat app Slack.

[4]https://slack.com
[5]https://api.slack.com/incoming-webhooks

notification. Slack users are able to click on links in their Slack clients to view the comment in WebProtégé (right inset in Figure 2).

In addition to comment threads, which can be used for adding review comments to a project, WebProtégé also allows entities to be "tagged" with tags that can be defined for a project. Figure 1 shows various tags in the class hierarchy that have been added as part of the review process. Tags and comments usually go hand-in-hand. Entities with specific tags can also be searched for in the search dialog. Entities can be automatically tagged according to rules, which users build using the same criteria that are available for searching entities (presented in Section 2.4). For example, entities that do not have a value for the rdfs:label annotation property could be automatically tagged with a tag Missing Label.

2.3 Change Tracking and Project History

One of the key features of WebProtégé is that all ontology changes are tracked and grouped into revisions. Revisions are based on atomic user interface operations, and it is possible to revert a single revision. Figure 3 shows an example of the history of a project. Changes can be viewed/filtered by entity, so that the entire history for an entity description (both logical and non-logical) can be examined. Labels for revisions are automatically generated based on the changes that occurred. In the full-blown OWL 2 editing interface it is possible to group a set of changes into a commit, which can then have a commit message assigned to it. WebProtégé allows downloading the set of ontologies in a given revision.

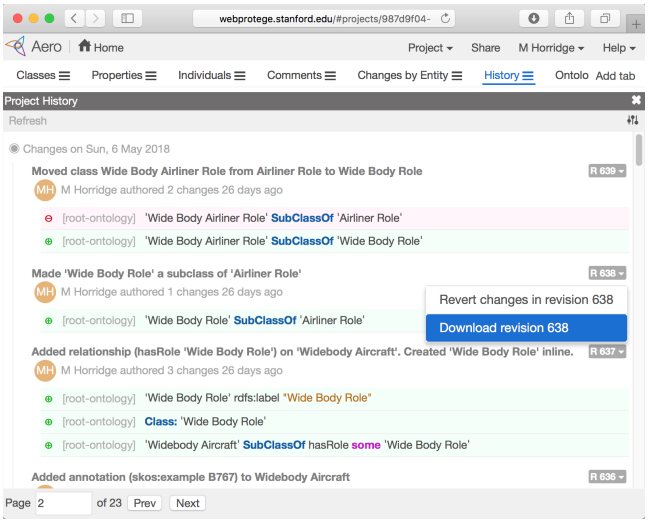


Figure 3: Change Tracking in WebProtégé. All ontology changes are tracked and grouped into revisions that can be reverted and downloaded. Revisions are labelled with a description, author and timestamp.

2.4 Querying

WebProtégé provides a simple user interface for performing complex ontology queries, shown in Figure 4. The kinds of queries supported essentially cover common search criteria and the basic expressivity of the SPARQL[6] query language.

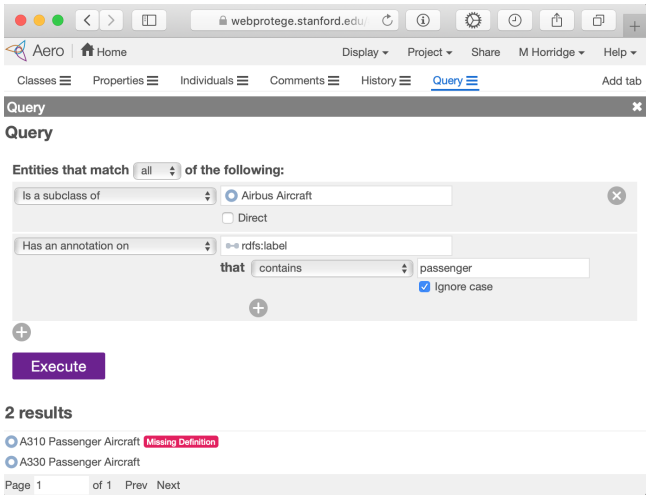


Figure 4: Querying in WebProtégé. Users can build arbitrarily complex queries with commonly used search criteria. The figure exemplifies a query for entities that match *all* criteria: that entities are subclasses of Airbus Aircraft, and that entities have a value for the rdfs:label annotation property containing the string "passenger", regardless of text case.

2.5 Graph Visualization

WebProtégé 4.0 features a brand new graph visualization, which we show in Figure 5. This visualization allows users to see all the relations that a selected entity takes part in. The unlabelled relations in the visualization represent rdfs:subClassOf relations, while labelled relations represent user-created properties. The unlabelled relations with a dashed line represent rdf:type relations. Using the graph visualization, users can isolate the relations (or "paths") in the graph between the selected entity and any entity that a user hovers over in the graph. Nodes in the graph can also be hidden from the visualization. WebProtégé allows users to download vector graphics files for the visualizations that it generates.

2.6 Integration with Third Party Applications

WebProtégé does not have a traditional "drop-in" plug-in mechanism like the desktop version of Protégé does. We intend to facilitate integration with third-party applications via Webhooks and a Web-API. As of now, third-party applications can be notified of the changes to a project by adding an event Webhook to a project. JSON messages are posted to the URL associated with the Webhook when changes in the project occur. We expect that this functionality will be used for loose coupling with third-party applications in the same way that GitHub and Slack can integrate with other applications.

[6] https://www.w3.org/TR/rdf-sparql-query

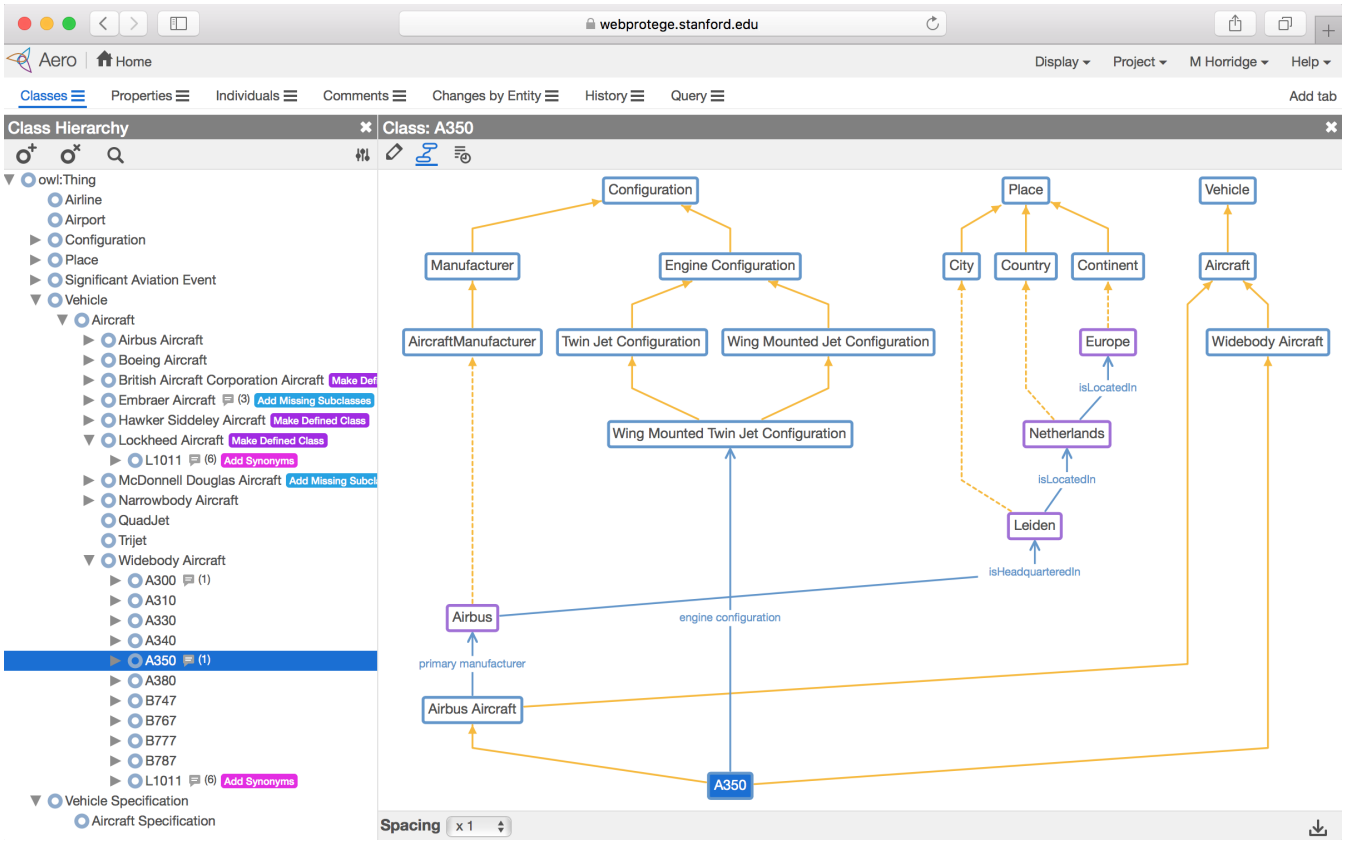


Figure 5: WebProtégé graph visualization. This visualization is focused on the class A350, and shows rdfs:subClassOf relations in yellow, rdf:type relations in dashed yellow, and custom relations in blue, between A350 and other classes such as Configuration.

3 SUMMARY

WebProtégé has been widely used since its inception. Important projects such as the 11th revision of the International Classification of Diseases (ICD-11) by the World Health Organization (WHO) depend on our software. The development environment used to edit ICD-11 is based on WebProtégé [4, 6]. We use WebProtégé to develop ontologies used in enterprise knowledge graphs. Some features such as search were inspired by these experiences.

The WebProtégé project would not be as successful as it is today without the engagement and the support of our massive user community. Our users help each other in the support mailing lists, and many of them interact closely with us on GitHub—by posting feature requests, issues, and prospective bug fixes or enhancements directly in our code repositories. We have also been fortunate to receive in-person feedback from our users as they learn how to develop ontologies using WebProtégé in our courses at Stanford University. Such feedback has proven to be invaluable in driving subsequent versions of the software.

In this paper, we have presented the major new features of the WebProtégé 4.0 ontology editor, developed at the Stanford Center for Biomedical Informatics Research. WebProtégé is under active development, and we seek feedback from the knowledge graph community so that we may continue to enhance WebProtégé to meet the needs of this community.

ACKNOWLEDGEMENTS

This work was funded by Grant GM121724 from the National Institute of General Medical Sciences at the United States National Institutes of Health.

REFERENCES

[1] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. 2008. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web* 6, 4 (2008), 309–322.

[2] Matthew Horridge, Tania Tudorache, Csongor Nyulas, Jennifer Vendetti, Natalya F Noy, and Mark A Musen. 2014. WebProtégé: a collaborative Web-based platform for editing biomedical ontologies. *Bioinformatics* 30, 16 (2014), 2384–2385.

[3] Mark A Musen. 2015. The Protégé project: a look back and a look forward. *AI Matters* 1, 4 (2015), 4–12.

[4] Tania Tudorache, Sean Falconer, Csongor Nyulas, Natalya F Noy, and Mark A Musen. 2010. Will semantic web technologies work for the development of ICD-11?. In *Proceedings of the International Semantic Web Conference*. Springer, 257–272.

[5] Tania Tudorache, Csongor Nyulas, Natalya F Noy, and Mark A Musen. 2013. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic Web Journal* 4, 1 (2013), 89–99.

[6] Tania Tudorache, Csongor I Nyulas, Natalya F Noy, and Mark A Musen. 2013. Using semantic web in ICD-11: three years down the road. In *Proceedings of the International Semantic Web Conference*. Springer, 195–211.

Content based News Recommendation via Shortest Entity Distance over Knowledge Graphs

Kevin Joseph
York University
Toronto, ON, Canada
KevinJoseph774@gmail.com

Hui Jiang
York University
Toronto, ON, Canada
hj@cse.yorku.ca

ABSTRACT

Content-based news recommendation systems need to recommend news articles based on the topics and content of articles without using user specific information. Many news articles describe the occurrence of specific events and named entities including people, places or objects. In this paper, we propose a graph traversal algorithm as well as a novel weighting scheme for cold-start content based news recommendation utilizing these named entities. Seeking to create a higher degree of user-specific relevance, our algorithm computes the shortest distance between named entities, across news articles, over a large knowledge graph. Moreover, we have created a new human annotated data set for evaluating content based news recommendation systems. Experimental results show our method is suitable to tackle the hard cold-start problem and it produces stronger Pearson correlation to human similarity scores than other cold-start methods. Our method is also complementary and a combination with the conventional cold-start recommendation methods may yield significant performance gains. The dataset, CNRec, is available at: https://github.com/kevinj22/CNRec

CCS CONCEPTS

• **Information systems** → *Structured text search*; • **Theory of computation** → *Shortest paths*; • **Applied computing** → *Document analysis*.

KEYWORDS

Content; Recommendation; News; Dataset

ACM Reference Format:
Kevin Joseph and Hui Jiang. 2019. Content based News Recommendation via Shortest Entity Distance over Knowledge Graphs. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317703

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317703

1 INTRODUCTION

News recommendation is a pivotal part of many major web portals and it is important for these search portals to recommend the "right" news articles, at the right time. Recommending news articles is one of the most challenging recommendation tasks because news providers generally allow users to read news articles anonymously. As a consequence, most news recommendation systems have to cope with the hard "cold start" problem: making recommendations without explicit user feedback.

A major difficulty in content-based news recommendation is that the similarity between news articles does not necessarily reflect their relatedness. For instance, two news articles might share a majority of words, yet their actual topic could be very different. Additionally one user's opinion of what is a good recommendation may significantly differ from another's. Over-specialization is also a problem with news recommendation [21]. In the news domain, many articles on the same topic are typically written differently on separate news portals. These articles, if recommended to the same reader, may result in a poor recommendation. Furthermore, the unstructured format of news articles makes it more difficult to automatically analyze their content than structured properties such as goods or services.

A common element across news stories is the description of specific events and/or updates about particular entities including people, places or objects. As these entities already represent key data related to the article, any additional information extracted about them could be used to make more informed recommendations. For instance, two news articles might share many common words, but the extra information gained through a knowledge graph may help to differentiate them. Through named entity disambiguation (NED), we determine the proper entity link from an optimized knowledge graph. Based on these entity links, a sub-graph can be created for each article, as shown in Figure 1. These articles are effectively embedded in the same space over the knowledge graph.

If the sub-graphs do not completely overlap, we assume these two articles mention related topics, but not the exactly same ones. In this case, it may be a good idea to make a content-based recommendation based on their distance over the knowledge graph. To do this we propose a new content-based news recommendation algorithm, which leverages the vast amount of information between these sub-graphs from a curated knowledge graph. This algorithm, similar to [28],

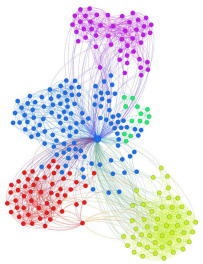


Figure 1: Each news article is represented as a subgraph (with different colours) in a large knowledge graph via named entity disambiguation.

computes the single source shortest path distance (SSSP), between two sub-graphs. Our experimental results have shown our proposed method produces stronger Pearson correlation to human similarity scores than other cold-start recommendation methods. Our method is also complementary to these conventional cold-start recommendation methods since an ensemble of them may yield significant performance gains. As the knowledge graph used in this work, Freebase [2], is extremely large, we focus our work on attaining the best results using the least data.

Unfortunately the similarity of two documents does not necessarily indicate that they would be a good recommendation to a user. Thus bench-marking using the LeePincombe-Welsh (LG50) [15] similarity dataset such as in [25], [24], and [28] would not be a good indicator of recommendation performance.

To get direct feedback, live recommendation tests as in [10] can be used, although this is time consuming and expensive. Live tests also exhibit time constraints that may not allow parameter tuning with algorithms requiring pre training or pre computations. For the reasons above we believe it is necessary to provide a new human annotated data set, *CNRec*. *CNRec* provides human annotated document-wise similarity scores as well as whether a pair of articles is considered as a good recommendation by human. This paper will also cover the creation of *CNRec* [1]. All results in this paper utilize *CNRec*. We hope *CNRec* will be used as a benchmark task for evaluating various cold-start news recommendation tasks in the research community.

2 RELATED WORK

Here we will briefly review related work on two topics: graph-based document similarity and content based news recommendation.

2.1 Graph-Based Document Similarity

PathSim in [30] defines a path between two sub-graphes A and B as the connectivity via the number of paths from A to B divided by the total number of path instances from A to B and B to A. *HeteSim* in [29] computes the relatedness between A and B as the overlap between the out-neighbors of A and the in-neighbors of B. Leal et al. [14] use a proximity metric as a measure of similarity among nodes, where proximity is defined as the state of being near in space, time, or relationship. Lam et al. [11] use a TF-IDF-inspired edge weighting scheme, combined with Markov centrality, to compute entity similarity. Palma et al. [22] describe an annotation similarity measure, called *AnnSim*, to evaluate similarity using a 1-1 maximal matching of annotations. *SemStim* algorithm [7] is based on PageRank and completes many random walks over all neighbors of a nodes.

Nunes et al. [20] present a document similarity approach where a document-wise connectivity score is computed based on the number of paths between document annotations. In a follow-up paper ,Nunes et al. use traditional TF-IDF to select top entities for each document. In [23, 24], Passant also computes all paths between two nodes and the number of direct and distinct links between resources in a graph, which are used to determine the similarity of two entities for recommendation.

Zhu *et. al.* [35] propose to use *WordNet* [18] and *DBpedia* [1] to determine the semantic similarity between concepts. They have attained the state of the art performance using a weighted shortest distance metric based on the least common subsumer measure extracted from graphs. This method measures similarity between pairs of nodes in a graph. In comparison, our work compares news articles by groups of nodes, i.e. sub-graphs.

In [28], similarity is measured based on entity linking and analysis of entity neighborhood in *DBpedia*, where information content is used to weight edges to compute the SSSP between all pairs entities in documents. In this work, we use a much larger knowledge graph, i.e., Freebase [2], and compute the average minimum symmetric distance across all pairwise entities between two articles.

2.2 Content-based News Recommendation

Term frequency inverse document frequency (*TF-IDF*) is a simple and popular method originally proposed for information retrieval [26]. TF-IDF is also widely used to determine the relative frequency of words in documents for content-based recommendation [21].

Doc2vec in [13] is an extension to learn an embedding vector for each document, where the order of words in the document is ignored and the objective is to predict a context word given the concatenated document and word vectors [12]. The learned document embedding vectors can be simply used to compute similarity between documents for content-based recommendation.

Bayesian Networks is a popular algorithm to model user interests in news recommendation [34], [32]. Context trees for news recommendation are proposed in [5] LDA is used in [16] to represent topic distributions [21]. [3] proposes locality

[1] CNRec is available here: https://github.com/kevinj22/CNRec.

sensitive hashing (LSH) and [16] uses *MinHash*, i.e., a probabilistic clustering method. *MinHash* has also been combined with probabilistic latent semantic indexing (pLSI) in [3].

In [31], a pre-trained named entity recognition (NER)model is applied to each news article. Based on the identified entities from articles, they determine user's long-term preferences for collaborative filtering. In this paper, rather than building user profile, we also leverage named entity information to improve content based recommendation rather than collaborative filtering. [16], [31] and [21] survey more natural language processing techniques used for news recommendation.

2.3 Collaborative Filtering

In collaborative filtering, recommendations are done by using the other preferences which are similar to the current users' past preferences. As this work focuses solely on content based recommendations this will not be covered here.

3 DATA SET: *CNREC*

CNRec provides document to document similarity as well as whether a pair of articles was considered a good recommendation. The *CNRec* data set consists of 2700 pairs of news articles, selected from 30 groupings of 10 articles of human perceived similarity 3. In total we have 300 unique news articles originally published in a period of 3 consecutive days between August 25-28, 2014. Each article is paired with all other articles in the same group. This results in 45 pairs that should produce positive similarity ratings. Another 45 pairs are randomly generated across other groups resulting in 2700 total pairs. The 3 day period, as well as the grouping and pairing procedure, provides a ideal set of articles to process. It allows engineers to focus on direct algorithm design rather than filtering relevant articles by time, or overlapping entities [25], before computations.

Each pair of articles is rated by 6 human annotators against two questions:

(1) In terms of content delivered, how similar do you think these two articles are? The annoators were given 3 choices: *Not Similar / Similar/ Very Similar*. Their answers were converted into numerical values 0/1/2 .
(2) If one of these articles was recommended based on the other would you have followed the link? Each annotator choose between *NO* and *YES*, which were converted to numerical values 0/1.

Referencing Figure 2, *CnRec* indicates there is a 21.42%, 43.40%, 35.18% split between similarity ratings leading to a good recommendation and a 73.71 Pearson correlation between similarity rating and good recommendations. This indicates that a high similarity does not mean a good recommendation.

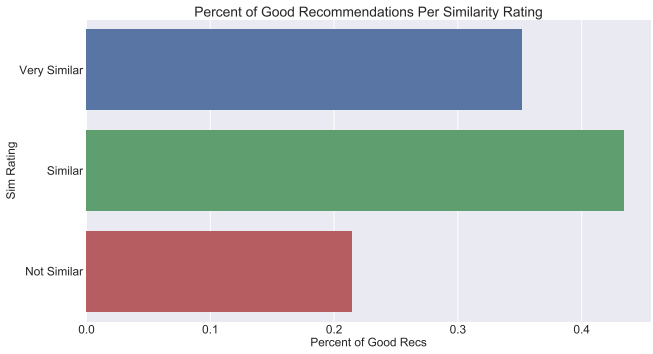


Figure 2: Similarity Rating Percent of Good Recommendations

(1) D'oh! Homer Simpson takes ALS Ice Bucket Challenge
(2) McHendry Local pastor takes part in Ice Bucket Challenge
(3) Pub manager to have ice cold water poured over him tonight
(4) Woman has different take on ALS Ice Bucket Challenge
(5) Let me tell you about my hero, the man who beat ALS
(6) Facebook user claims that only 7% of ALS challenge donations go towards charity
(7) Anti-Abortion Activists Have a Big Problem With the Ice Bucket Challenge
(8) ALS Ice Bucket Challenge Charity challenge hits Chelmsford
(9) Downton Abbey's Michelle Dockery's completes the ALS Ice Bucket challenge
(10) George RR Martin does Ice Bucket Challenge (then slips in the pool)

Figure 3: Example CNRec Similar Group Article Titles

4 ENTITY SHORTEST DISTANCE OVER KNOWLEDGE GRAPHS

In this section, we introduce the main ideas of the method we propose for content-based news recommendation, namely entity shortest distance over knowledge graphs (SED).

4.1 Knowledge Graphs

Many public Knowledge Graphs (KGs) are currently available. These include Freebase, DBpedia, and YAGO. Each network records millions of concepts, entities and their relationships.

As news articles are centered on various named entities, one could hypothesize that by using a knowledge graph composed of said entities, a reasonable recommendation could be made. In this paper all experiments are conducted with the Freebase [2] KG. In Freebase all relationships are modeled in the Resource Description Framework (RDF) format [2], where each node edge pair is modeled in an object-orientated manner as object-predicate-subject. A leaf, a vertex of out degree 1, in Freebase is a human readable string used to assign values to an entities, such as birthdays or genders. The

edges in Freebase are labeled by a predicate which exhibits a *domain.type.property* relationship between its related vertices [2]. Note that Freebase is originally modeled as a directed graph but we do not consider the direction of edges because most relations have a sound inverse relation, such as *nationalyOf* and *ofNationality*. This decision follows that of [28], [7], and [35].

Freebase is a huge knowledge graph consisting of approximately 1.9 billion triplets [2]. Any undirected simple graph G holds $|E| <= (|V| * (|V| - 1))/2$ edges. This implies $|V|^2$ quadratic growth in the number of edges as |V| increases. Thus the number of edge traversals required to find the shortest path grows exponentially along with path length. Therefore we optimize the original Freebase KG by pruning some redundant and uninformative vertices and edges. In particular, all non-English entries are removed (*CnRec* is English only). Many nodes in Freebase with out-degree less than 20 are generally filled with redundant information, thus they are also removed in our experiments. Nodes that appeared too frequency in our initial path searches and their associated edges were removed. For example *Male*, *people.person*, *olympic_participating_country*, and *United States*; these all fail to provide any specific information for recommendation. In this way, we may retain the most prominent information, remove noise, and reduce computational complexity. The English only Freebase KG is composed of 438 000 million triplets, while the completely pruned Freebase has 17.5 million triplets.

4.2 The Proposed SED Algorithm

Our proposed SED recommendation algorithm consists of the following steps:

(1) Entity discovery and linking: identify all named entities from each news article and link each entity to a node in Freebase
(2) Sub-graph generation: create a sub-graph from the knowledge graph based on the linked nodes
(3) Shortest distance similarity between articles: compute pair-wise shortest distance between any two documents over a union of their corresponding sub-graphs.

4.2.1 Entity Discovery and Linking (EDL): mapping news articles to KG. Given each news article we can use entity discovery and linking (EDL) to map it to an knowledge graph. EDL first identifies all named or nominal mentions of entities from unstructured text data. Next it classifies them onto proper types: person (PER), location (LOC), Organization (ORG), Geopolitical Entity (GPE) and Facility (FAC). Lastly, EDL needs to link each identified entity to a correct entity node in a knowledge graph. For example, in a sentence like "*Apple just released its new iPhone.*", "*Apple*" needs to be a node in Freebase associated with *Apple Inc.* rather than the fruit. In this work, we do not study how to improve EDL performance, but directly use an off-the-shelf EDL system [17, 19]. This system won the TAC KBP trilingual EDL competition in 2016 [9].

4.2.2 Sub-graph Creation. After EDL, each news article may be represented as a sub-graph in the shared knowledge graph. To produce a sub-graph, we start with a set of discovered entities N and create a labeled undirected graph $G = (V, E)$. First we set of nodes V of G be composed of all discovered entities: $V = N$. Next for each entity n in N we expand the nodes in G. This is done by performing a breadth first search on Freebase adding all outgoing relations. Loops are not added in graph creation. When used with our proposed weighting scheme all connections between node pairs are collapsed into one. All related predicates are then stored in an edge list. As our proposed weighting scheme considers the number of common neighbors, these paths would all be weighted equally. This removes the necessity of keeping all unique connections between nodes, reducing computational complexity. We expand up to a maximal length L for each discovered entity effectively adding all simple paths of up to length L. From this we attain a sub-graph of Freebase. This graph is composed of the discovered entities and their related concepts along all paths of maximal length L that connect them.

4.2.3 Shortest Distance Between Articles. The most intuitive information for similarity or relatedness between any two articles is the distance between their sub-graphs in the KG. Conceptually speaking, the closer these two sub-graphs are located to each other, the more similar these two articles may be.

Let $\mathcal{P}(n_i, n_j) = \{P_1, P_2, ..., P_k\}$ be the set of all possible paths connecting the nodes n_i and n_j with cardinality or size N. Let $|P_i|$ denote the length of a path $P_i \in \mathcal{P}(n_i, n_j)$, then $D(n_i, n_j) = \min(|P_k|)$ denotes the shortest path length between two nodes, namely shortest distance between them over G.

With EDL, a set of named entities is extracted from each article. After linking them to nodes in KG, each article is modeled by a subgraph in KG. As symmetric algorithms are provided in [30], [29], [27], and [25] and one can easily derive asymmetric version we only present the symmetric version of SED. We measure the similarity between two articles, e.g. S_1 and S_2, based on the shortest distance between their subgraphs over the KG, denoted $\mathcal{D}(S_1, S_2)$ as:

$$\hat{\mathcal{D}}(S_1, S_2) = \frac{\mathcal{D}(S_1, S_2) + \mathcal{D}(S_2, S_1)}{2}$$

where $\mathcal{D}(S_1, S_2) = \frac{\sum_{m \in S_1} \min_{n \in S_2} D(m,n)}{|S_1|}$.

Note that this method provides an average minimum row-wise distance. This can easily be modified to use the average or sum of distance across all all pairwise nodes between two articles. By using the minimum we focus on the best pair of entities which has a higher probability of cooccurring in documents. This follows [25] which computes an average maximum symmetric similarity between two articles. If we impose weights onto all edges in KG, we may compute weighted shortest distance as the sum of all weights along the traversed edges.

5 CREATING THE OPTIMAL SUBGRAPH

There are many practical issues to be considered in creating the subgraph for each news article. In this section, we will discuss some of these practical issues.

5.1 Maximal Length Expansion L

Does a larger graph have a significant impact on performance? Following a large body of evidence from work [8], [28], [30] and [25] we only consider 1 and 2 hop graph creation. Additionally by using the union of two articles subgraphs, these 1 and 2 hops per node, which have been noted to have high predictive power, allow for longer relevant path lengths between the articles.

5.2 Pre-Screening Entities

Assume a set of named entities is identified by the EDL system [33] for an article, how do we know whether some entities are more important than others or to treat them all equally? One may perceive that the more an entity appears in an article, and the earlier it appears, the more closely related it is to the main theme of the article. Combining these two statistics, we rank all entities by a stable sort: first in descending order by count then in ascending order by offset. Over *CnRec* 11 entities are found on average per article. In addition only a single article returns 0 entities. The EDL system also returns an entity type for each identified entity, such as Person (PER), Location (LOC), Organization (ORG), Geopolitical Entity (GPE) and Facility (FAC). Some types have many more neighbouring nodes in the KG than others, such as GPE as shown in Figure 4. Others such as LOC are frequently mentioned in articles. To insure there is no over saturation of common entities we can filter out entities. For our experiments we remove LOC and GPE typed entities. This leaves 7.5 entities on average per article. We then try filtering the top 5 and 8 entities per article.

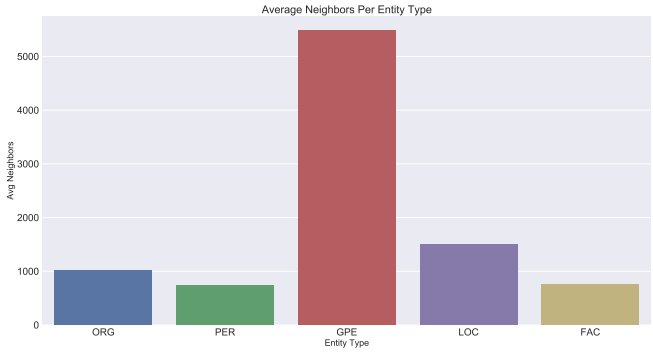


Figure 4: Average Entities Per Type

5.3 Augmenting Context Words

In this work, we use an off-the-shelf EDL system,[9] that is not designed to capture all Freebase entities. Moreover, some keywords in a news article may not be an above-defined entity at all. For example, in an article about Ebola, a debilitating virus, the word *Ebola* may be a key concept to find related articles. *Ebola* is not a named entity by definition. To address these issues, we propose to augment some "context words" from each news article. We use TF-IDF to score words for each article taking the top scoring N words which exactly match a Freebase node's title. These N words are added to the article's sub-graph. This method insures no articles have 0 discovered entities. In this paper, we experiment with $N = 2-4$ context words per article.

5.4 Penalty for Disconnected Entities

If there is no path between two entities across a pair of articles setting a 'penalty' value could impact the recommendation quality. In the event the penalty value is set too low, too many articles would be recommended. If it is set to high, not enough articles will be recommended. In this paper we experiment within a range of penalty values.

6 WEIGHTING GRAPH EDGES

In this work, we propose to weigh each Freebase edge based on social network influence weighting [6], which is named *relation weighting scheme (RWS)* hereafter.

6.1 Relation Weighting Scheme

The proposed relation weighting scheme (RWS) is based on social network influence weighting [6]. Given a sub-graph G for an article, assume we have an edge, e, which is connected to two entity nodes, i.e., n_i and n_j. For each entity, we may query the KG and obtain all first-order neighbouring nodes of each to form two neighborhood sets, denoted as $\mathcal{N}_i$ and $\mathcal{N}_j$. To weigh the edge, e, between n_i and n_j, we first compute the conditional probability as follows:

$$p(n_j|n_i) = \frac{|\mathcal{N}_i \cap \mathcal{N}_j|}{|\mathcal{N}_i|}.$$

If two nodes share many neighbours in common it is likely that they are closely related. As we are computing the shortest distance, a lower edge weight is better. As we are querying a pair of neighboring nodes $p(n_j|n_i) > 0$, then a weight, w_i, is assigned to e as: $w_i = 1 - p(n_j|n_i)$.

6.2 Other Weighting Schemes

In [28], and [4] many metrics for ranking edges in a KG are proposed. In this work, we implement JointIC and [4]'s top three metrics to weight edges for our news recommendation tasks: i) AF: normalized Attribute Frequency of the attributes over KG, favors common attributes in KG; ii) IAF: inverse attribute frequency for the attributes over KG, favors rare attributes; iii) AF-IAF: multiply the previous two features.

7 EXPERIMENTS

In this section, we evaluate the recommendation performance of the proposed methods. These are evaluated on a newly-created news recommendation benchmark data set, *CNRec*, on some typical cold-start settings. The proposed method

is compared with conventional approaches used for cold-start content-based news recommendation, including TF-IDF and doc-vector. It is also compared with other graph traversal algorithms in terms of F_1, precision, recall, Pearson, and Spearman correlation with human labeled scores. All experiments with SED use the pruned version of FreeBase in which we removed nodes with out degree less than 20, symmetric distance, the top 5 non LOC / GPE entities, leaves, and are weighted unless stated otherwise.

Table 1: Positive Article Pairs for Evaluation Conditions (%)

GR@.75	GR@.5	DR@.75	DR@.5
25	40	8	21

7.1 Evaluation Conditions

The two evaluation questions are highly subjective and the answers varied on an individual basis. For example, for question 2 only 15% of article pairs were unanimously considered to be a good recommendation by all six participants. Therefore, in this work, we will evaluate our recommendation algorithms under the following four different conditions. In each condition, the *CNRec* data set is split into positive and negative samples in a different way to take the subjectivity issue into account.

(1) Good Recommendation at least XX% (*GR@.50* and *GR@.75*): A pair of articles is considered as a good recommendation if the average rating score of Question 2 across all participants is 0.XX or above.
(2) Diverse Recommendation at least XX% (*DR@.50* and *DR@.75*): A pair of articles is considered as a good recommendation only if at least XX% of participants think it is a good recommendation (Yes to Question 2) and 50% of participants think they are not very similar (Question 1 rating is 1 or below).

From Table 1 we can see that *GR@.50* will produce 40% positive pairs. For *GR@.75* 25% of article pairs are positive samples. Using *DR@.50* we have 21% positive samples. Lastly for *DR@.75* we have 8% as positive samples.

In our experiments each algorithm's raw recommendation scores are normalized to be zero mean and unit variance. The scores are then judged against a universal threshold, 0 in this case, to make a binary recommendation decision. For each of the above conditions, the recommendation results are compared with the positive/negative labels to calculate precision and recall. The final recommendation performance is measured by F_1 score, i.e., $F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$.

7.2 Analysis on various hyperparameters

In this section, we will study how the choices of some critical hyper-parameters in the proposed algorithm may affect the recommendation performance.

7.2.1 Optimal Maximal Length Expansion Radius L. From 2 we can see that the one hop weighted subgraphs excel in all categories. The weighting is much more effective for the 1 hop graphs than the two hop graphs. We believe this is because our RW checks the immediate neighbors of two connecting nodes.

Table 2: Maximal Length Expansion Radius (F_1 scores in %). (Penalty 0.98 was used in all tests. No LOC or GPE entities were used.)

Weighted / Hops	GR@.75	GR@.5	DR@.75	DR@.5
UnW / 2	64.57	70.23	23.16	38.26
W / 2	64.87	71.62	23.79	39.75
UnW / 1	64.45	72.07	23.51	40.37
W / 1	**65.09**	**73.54**	**24.72**	**42.49**

Table 3: Entity pre-screening and context words performance comparison (F_1 scores in %). (All 1 Hop; W: weighting; NE: Number of Entities where NLG indicates No LOC or GPE. C: adding # context words; Penalty 0.98 was used in all tests.)

W / NE / C	GR@.75	GR@.5	DR@.75	DR@.5
UnW / All (11 Avg) / 0	60.79	67.58	21.37	36.89
UnW / NLG 8 / 0	64.45	72.07	23.51	40.37
UnW / NLG 5 / 0	72.53	78.46	27.81	44.98
UnW / NLG 5 / 2	**75.89**	**83.76**	**29.41**	49.47
UnW / NLG 5 / 4	74.30	84.25	28.64	**49.94**
W / All / 0	64.45	71.02	23.88	39.57
W / NLG 8 / 0	65.09	73.54	24.72	42.49
W / NLG 5 / 0	72.79	80.02	28.42	46.84
W / NLG 5 / 2	**75.91**	**84.72**	**29.41**	49.76
W / NLG 5 / 4	74.29	84.67	28.82	**50.66**

7.2.2 Effect of entity screening and context words. As shown in Table 3, after entity pre-screening and removing all entities of the type of LOC and GPE, we observe large gains across four evaluation conditions for both unweighted and weighted distance calculation. It is observed that while common nodes of LOC and GPE types may be useful in some situations, they contain too many neighboring nodes. This in turn, leads to many incorrect recommendations due to common entity overlapping. After removal of LOC and GPE there were 7.5 average nodes per article. Thus we skipped filtering top 10 entities in favor of top 8 and 5. Furthermore, we can see it is helpful to add a small number of context words for each article. Based on the results in Table 3, we have decided to add up to 2 context words to each article since this yields the largest gain across various categories.

7.2.3 Effect of various edge weighting schemas. Here we compare our proposed edge weighting method, RWS, in section 6.1, with four existing methods, i.e., AF, IAF and AF-IAF, JointIC, with SED. As shown in Figure 5, we can see that RWS produces much larger gains across all evaluation conditions over the three existing weighting methods. We believe this is because RWS weights each node based on the underlying article while the three traditional methods are purely based on term frequency over the entire knowledge graph.

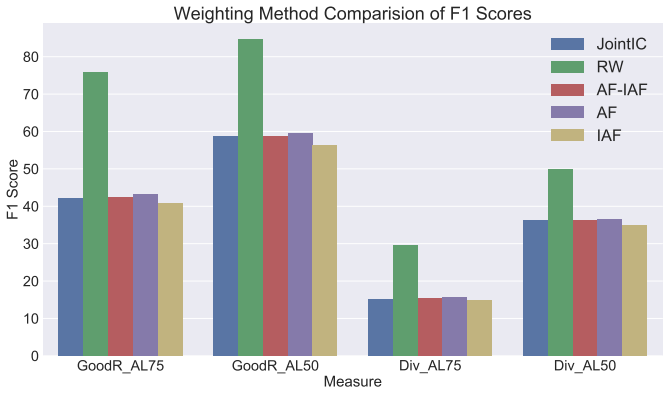


Figure 5: Edge weighting method comparison via Sym SED, entity pre-screening, 2 context words.

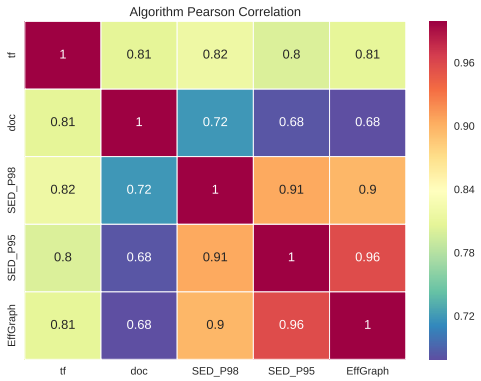


Figure 6: Pearson correlation of recommendation scores from various algorithms.

Table 4: SED Penalty Values: F1 Scores (%)

	GR@75	GR@50	DR@75	DR@5
NoP	74.19	81.70	29.35	48.47
P98	**75.91**	**84.72**	**29.41**	**49.76**
P95	74.24	81.43	29.40	48.01
P90	70.53	77.44	26.03	44.39

All tests use symmetric distance, filtered entities, 2 context words, and relational weighting. PXX indicates a penalty of 0.XX was used.

These methods likely impart less specific information on the active article.

7.2.4 Penalty for Disconnected Nodes. Our experiments from table 4 show by setting a lower penalty value, the overall accuracy of the algorithm is decreased. We believe this is because too many unrelated articles are recommended.

7.3 Computation Time

All graphs created are of expansion radius 1 which provided the best results for all algorithms. Using our filtered Freebase of 17.5 million nodes graphs were created in 50 milli-seconds (ms) on average. Weighting took 78.8 ms average per graph. The English only Freebase of 438 000 million nodes took 512 ms on average to create each graph. As EffGraph (EG) performed better using the unfiltered English only Freebase, it was used to create it's graphs. The average time for the 689 entity graph creation was 35 ms. Computing the all pairwise shortest distance to all other discovered entities (entities were skipped if not present in the graph) took 13.2 ms on average, although a large majority of entity pairs (99.26%) were skipped. The 5400 evaluations took 29.4 ms on average for SED on the 17.5 million Freebase, 983 ms for the 438 000 SED graphs, 30.94 for KGD, and 0.00017 for EG. Computations used a Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz, with 8 GB of RAM.

7.4 Baseline Parameters

For experiments utilizing TF-IDF all stop words are removed from the corpus. We then filter out any words in over 80% of articles. The Doc2Vec model is the distributed memory (PV-DM) model. Past research has shown that PV-DM outperforms the the distributed bag of words (PV-DBOW) model when the corpus is small. As our data set is 300 articles this is a reasonable choice. Our embedding size is set to 300.

As SED computes the distance between two articles TF-IDF and doc2Vec were converted from similarity to distance measures. Since both similarity computations were done with cosine similarity they output values between -1 and 1. To bound the results between 0 and 1 any values that were less than 0, indicating that the vectors are diametrically opposed, were set to 0. Then to convert from similarity to distance we subtracted the similarity value from 1.

7.5 Main Results

In Table 5, we show the performance of two baseline recommendation algorithms *TF-IDF* and *Doc2vec*, in terms of F_1 scores for all four evaluation conditions. From there, we see that *TF-IDF* outperforms *Doc2vec* by a large margin in *GR@.75* and *DR@.75* while *Doc2vec* displays higher accuracies in the at-least-50% tests.

Next, we have evaluated several configurations of the proposed shortest entity distance (SED) method for the four evaluation conditions. From the results in Table 5, it is clear that the [28] bipartite matching and all-average distance calculation performs the worst in all cases. This may be due to the fact that the distance measures from some critical entities are over-averaged or ignored due to prior matchings by other less importance entities. The symmetric distance computation yields the best performance across all categories. We believe this is because the symmetry prevents one article with many entities from encapsulating the other.

By itself *SED* symmetric performs much better than *Doc2Vec* in *GR@.75* and *DR@.75* categories. It gives higher results versus *EG* in both *@ 0.5* categories. It also provides better

Table 5: Recommendation Algorithm Performance.

		Conventional		Graph Based				
	Algo	TF-IDF	Doc2Vec	EG	KGD	SED Avg P98	SED Row P98	SED Sym P98
DR@50	F1	48.66	**55.76**	48.80	41.01	44.38	48.73	**49.76**
	Precision	63.29	**91.43**	63.81	58.74	55.59	64.60	**67.31**
	Recall	39.52	**40.11**	**39.50**	31.50	36.93	39.12	39.47
DR@75	F1	**29.28**	27.19	**30.31**	24.89	26.03	29.25	29.41
	Precision	78.20	**97.63**	81.52	75.36	66.11	80.09	**82.70**
	Recall	**18.01**	15.80	**18.61**	14.91	16.20	17.89	17.89
GR@50	F1	85.56	**85.89**	83.81	73.68	77.43	83.21	**84.72**
	Precision	79.37	**95.24**	78.06	73.53	69.84	78.29	**80.95**
	Recall	**92.79**	78.22	**90.48**	73.84	86.88	88.78	88.88
GR@75	F1	**78.42**	68.04	**77.04**	66.82	70.53	75.27	75.91
	Precision	91.80	**98.98**	90.63	85.58	79.72	89.68	**92.17**
	Recall	**68.45**	51.84	**66.99**	54.81	63.24	64.85	64.53

Table 6: Recommendation Algorithm Ensembles Performance.

	Algo	EG Doc	EG Doc TF	EG TF	SED Doc	SED Doc TF	SED TF	TF Doc
DR@50	F1	56.48	55.73	47.93	**57.19**	56.17	49.03	56.49
	Precision	**88.99**	78.41	61.71	88.29	79.81	63.81	81.03
	Recall	41.37	43.23	39.18	42.29	43.33	39.80	**43.36**
DR@75	F1	28.37	**31.48**	30.04	29.25	31.47	30.59	30.86
	Precision	96.92	93.13	79.15	**97.39**	94.31	81.75	93.60
	Recall	16.62	**18.94**	18.53	17.21	18.89	18.81	18.48
GR@50	F1	87.42	89.45	85.40	88.39	**89.62**	86.02	89.49
	Precision	**93.93**	88.05	78.62	93.46	88.89	79.83	89.40
	Recall	81.76	90.89	**93.45**	83.84	90.37	93.24	89.57
GR@75	F1	70.50	77.19	79.42	71.87	76.88	**79.44**	75.86
	Precision	98.76	97.22	92.09	**98.76**	97.73	93.05	97.29
	Recall	54.82	64.00	**69.81**	56.49	63.36	69.30	62.16

performance in the *DR@.75* case over *TF-IDF* and *Doc2Vec* although lower than *EffGraph*. In the *DR@.50* category, the SED symmetric variant outperforms *TF-IDF* and *EG* but is far behind *doc2vec*. From these observations, it may be concluded that SED leverages additional information from both the context and entity nodes to make more refined diverse recommendations than the baseline algorithms, but less refined versus *EG*.

As shown in Figure 6 we can see that the SED correlates less with *Doc2vec* and *TF-IDF* than they do with each other. The results in Table 5 show that a consolidation of SED, TF-IDF, and doc2vec provides the highest performance in *GR@.75* producing a gain of 1.02% over the blending of *TF-IDF* and *doc2vec*. A combination of SED and *doc2vec* provides the highest overall performance in *GR@.5* producing a gain of 0.13% over the second best of *TF-IDF* and *doc2vec* . A synthesis of SED and *TF-IDF* provides the highest performance in *DR@.75* producing a gain of 0.62% over the second best of *TF-IDF* and *doc2vec*. A combination of SED and *doc2vec* give the best performance in *DR@.5*. This accounts for a gain of 0.70% over the second best of *TF-IDF* and *doc2vec*. Our hypothesis is that the smaller correlation between SED and the baseline algorithms accounts for the gains as SED would recommend articles the baseline algorithms would not.

Finally, we evaluate the correlation between the algorithm's recommendation scores and the average annotation scores of Question 2 among the six human annotators. In Table 7, we first list the correlation coefficients (both Pearson and Spearman) of the baseline algorithms: *TF-IDF* and *Doc2vec*. As we can see, the correlation coefficients of SED alone is much lower than both baseline algorithms. Only 5.3% of distance values returned a 1 when using SED. While many other values returned close to 1, these values do no correlate to the human similarity scores that were exactly 1. When using EG 63.1% of distance values were 1. We believe this is why SED has higher precision VS EG in all categories and produces a higher Pearson correlation versus both other traversal algorithms, but lower Spearman VS EG.

8 CONCLUSIONS

The proposed shortest entity distance (SED) algorithm performs the best under almost all examined conditions in the cold-start content-based news recommendation task when it is configured with maximal length 1 expansion, relation

Table 7: Correlation coefficients between average human annotation scores of Question 2 and recommendation scores from various algorithms.

	Pearson	Spearman
Doc2Vec	78.832	81.155
TF-IDF	**87.709**	**81.469**
KGD	70.093	61.413
EG	78.975	**78.870**
SED	**79.273**	72.730
TF + Doc	88.310	83.456
SED + TF	88.562	81.899
SED + Doc	84.271	82.642
SED + Doc + IF	**88.957**	**83.852**

weighting scheme, entity pre-screening, leaf nodes, and up to 2 context words. Moreover, the above SED algorithm yields significant gains in these tasks when combined with some conventional recommendation algorithms. This has shown that the proposed SED algorithm has effectively leveraged useful information from knowledge graphs for content-based recommendation, which is fairly complementary with most existing algorithms. SED also provides the highest Pearson correlation of graph traversal algorithms.

Lastly, we have created a new data set for content-based news recommendation from this work. We hope that this data set will become a standard benchmark in the content-based news recommendation field as it provides a variety of human opinions and a diverse set of answers allowing researchers to evaluate their algorithms across many metrics.

REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference (ISWC'07/ASWC'07)*. Springer-Verlag, Berlin, Heidelberg, 722–735. http://dl.acm.org/citation.cfm?id=1785162.1785216

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*. ACM, New York, NY, USA, 1247–1250. https://doi.org/10.1145/1376616.1376746

[3] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 271–280. https://doi.org/10.1145/1242572.1242610

[4] A Caputo E Ali and S Lawless. 2017. Entity Attribute Ranking Using Learning to Rank. http://ceur-ws.org/Vol-1883/paper_10.pdf. *CEUR* 1883 (2017).

[5] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized News Recommendation with Context Trees. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 105–112. https://doi.org/10.1145/2507157.2507166

[6] Sudheendra Hangal, Diana L. MacLean, Monica S. Lam, and Jeffrey Heer. 2010. All Friends are Not Equal: Using Weights in Social Graphs to Improve Search. http://vis.stanford.edu/files/2010-WeightsSocialGraphs-SNAKDD.pdf.

[7] B. Heitmann and C. Hayes. 2016. SemStim: Exploiting Knowledge Graphs for Cross-Domain Recommendation. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. 999–1006. https://doi.org/10.1109/ICDMW.2016.0145

[8] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised Graph-based Topic Labelling Using Dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 465–474. https://doi.org/10.1145/2433396.2433454

[9] Heng Ji and Joel Nothman. 2016. Overview of TAC-KBP 2016 Tri-lingual EDL and Its Impact on End-to-End Cold-Start KBP. In *Proceedings of 2016 Text Analysis Conference (TAC)*. https://tac.nist.gov/publications/2016/papers.html

[10] Benjamin Kille, Andreas Lommatzsch, Frank Hopfgartner, Martha Larson, and Torben Brodt. 2017. CLEF 2017 NewsREEL Overview: Offline and Online Evaluation of Stream-based News Recommender Systems. http://eprints.gla.ac.uk/143586/

[11] Samantha Lam. 2013. Using the Structure of DBpedia for Exploratory Search. http://chbrown.github.io/kdd-2013-usb/workshops/MDS/doc/mds2013_submission_2.pdf.

[12] Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *CoRR* abs/1607.05368 (2016). http://arxiv.org/abs/1607.05368

[13] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). http://arxiv.org/abs/1405.4053

[14] JosÃľ Paulo Leal, VÃćnia Rodrigues, and Ricardo QueirÃąs. 2012. Computing Semantic Relatedness using DBPedia.. In *SLATE (OASICS)*, Alberto SimÃţes, Ricardo QueirÃąs, and Daniela Carneiro da Cruz (Eds.), Vol. 21. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 133–147. http://dblp.uni-trier.de/db/conf/slate/slate2012.html#LealRQ12

[15] Michael D. Lee and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. http://www.socsci.uci.edu/~mdlee/lee_pincombe_welsh_document.PDF. In *In CogSci2005*. Erlbaum, 1254–1259.

[16] Lei Li, Ding-Ding Wang, Shun-Zhi Zhu, and Tao Li. 2011. Personalized News Recommendation: A Review and an Experimental Investigation. *J. Comput. Sci. Technol.* 26, 5 (Sept. 2011), 754–766. https://doi.org/10.1007/s11390-011-0175-2

[17] Dan Liu, Wei Lin, Shiliang Zhang, Si Wei, and Hui Jiang. [n. d.]. Neural Networks Models for Entity Discovery and Linking. In *arXiv:1611.03558*. https://arxiv.org/abs/1611.03558

[18] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

[19] Mingbin, Feng Wei, Sedtawut Watcharawittayakul, Yuchen Kang, and Hui Jiang. 2016. The YorkNRM Systems for Trilingual EDL Tasks at TAC KBP 2016. In *Proceedings of 2016 Text Analysis Conference (TAC)*. https://tac.nist.gov/publications/2016/papers.html

[20] Bernardo Pereira Nunes, Ricardo Kawase, Besnik Fetahu, Stefan Dietze, Marco A. Casanova, and Diana Maynard. 2013. Interlinking Documents based on Semantic Graphs. *Procedia Computer Science* 22, Supplement C (2013), 231 – 240. https://doi.org/10.1016/j.procs.2013.09.099 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013.

[21] Ozlem Ozgobek, Jon Atle Gulla, and R. Cenk Erdur. 2014. A Survey on Challenges and Methods in News Recommendation. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies - Volume 2: WEBIST,*. INSTICC, SciTePress, 278–285. https://doi.org/10.5220/0004844202780285

[22] Guillermo Palma, Maria-Esther Vidal, Eric Haag, Louiqa Raschid, and Andreas Thor. 2013. Measuring Relatedness Between Scientific Entities in Annotation Datasets. https://bit.ly/2HAyeVA.

[23] Alexandre Passant. 2010. Dbrec: Music Recommendations Using DBpedia. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part II (ISWC'10)*. Springer-Verlag, Berlin, Heidelberg, 209–224. http://dl.acm.org/citation.cfm?id=1940334.1940349

[24] Alexandre Passant. 2010. Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. https://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1147.

[25] Christian Paul, Achim Rettinger, Aditya Mogadala, Craig A. Knoblock, and Pedro Szekely. 2016. Efficient Graph-Based Document Similarity. In *Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains - Volume 9678*. Springer-Verlag New York, Inc., New York, NY, USA, 334–349. https://doi.org/10.1007/978-3-319-34129-3_21

[26] Juan Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. https://bit.ly/2HRlE7J.

[27] A. Sanfeliu and K. S. Fu. 1983. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13, 3 (May 1983), 353–362. https://doi.org/10.1109/TSMC.1983.6313167

[28] Michael Schuhmacher and Simone Paolo Ponzetto. 2014. Knowledge-based Graph Document Modeling. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 543–552. https://doi.org/10.1145/2556195.2556250

[29] Chuan Shi, Xiangnan Kong, Yue Huang, Philip S. Yu, and Bin Wu. 2013. HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. *CoRR* abs/1309.7393 (2013). http://dblp.uni-trier.de/db/journals/corr/corr1309.html#ShiKHYW13

[30] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. https://bit.ly/2Hx6XTM. In *In VLDB 11*.

[31] Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C. Almeroth, Jon Espen Ingvaldesn, Gaute Nygreen, and Erik Berg. 2013. Tailored News in the Palm of Your Hand: A Multi-perspective Transparent Approach to News Recommendation. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13 Companion)*. ACM, New York, NY, USA, 305–308. https://doi.org/10.1145/2487788.2487930

[32] Hao Wen, Liping Fang, and Ling Guan. 2012. A Hybrid Approach for Personalized Recommendation of News on the Web. *Expert Syst. Appl.* 39, 5 (April 2012), 5806–5814. https://doi.org/10.1016/j.eswa.2011.11.087

[33] Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A Local Detection Approach for Named Entity Recognition and Mention Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1237–1247. http://aclweb.org/anthology/P17-1114

[34] Linda Yang and K. Yeung. 2010. A proactive personalized mobile news recommendation system.

[35] G. Zhu and C. A. Iglesias. 2017. Computing Semantic Similarity of Concepts in Knowledge Graphs. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (Jan 2017), 72–85. https://doi.org/10.1109/TKDE.2016.2610428

Building a Knowledge Graph for the Air Traffic Management Community

Richard M. Keller
Intelligent Systems Division
NASA Ames Research Center, Moffett Field, CA, USA rich.keller@nasa.gov

ABSTRACT

Historically, most of the focus in the knowledge graph community has been on the support for web, social network, or product search applications. This paper describes some of our experience in developing a large-scale applied knowledge graph for a more technical audience with more specialized information access and analysis needs – the air traffic management community. We describe ATMGRAPH, a knowledge graph created by integrating various sources of structured aviation data, provided in large part by US federal agencies. We review some of the practical challenges we faced in creating this knowledge graph.

CCS CONCEPTS

• Computing methodologies~Semantic networks • Information systems~Information integration

KEYWORDS

knowledge graph, ontology, semantic web, air traffic information management

ACM Reference format:

Richard M. Keller. 2019. Building a Knowledge Graph for the Air Traffic Management Community. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3317706

1 Introduction and Motivation

Every day, global aviation industry data providers generate a vast array of aviation information. When taken together, these data characterize the functioning of the global aviation system. The availability of this data provides the tantalizing possibility that we might analyze these data and discover new ways to improve air transportation economics, efficiency, and safety, for the benefit of all. However, aviation data are highly heterogeneous and are produced by a multitude of different providers in different formats and encodings. Improvements in the performance of the overall aviation system therefore depend on our ability to integrate, query, and analyze this information.

At the US National Aeronautics and Space Administration (NASA), we have been examining the use of knowledge graph technologies to create an integrated dataset for query and analysis of aviation data. Such a resource would be of potential interest to many aviation stakeholders, including public policy makers, airspace operators, flight controllers, airline carriers, aerospace researchers, and aviation industry service providers. These stakeholders require multiple sources of information to assist them in making decisions that impact current or future aviation systems operation.

We have focused our initial work on providing a resource for aerospace researchers who study air traffic management (ATM) procedures and systems. These researchers access a core set of ATM data sources generated by a handful of different providers, extract the data they require, and then write code to integrate the data – all before beginning their specific analyses. For researchers whose expertise lies in aeronautics – not data management – the overhead required to achieve data integration can be considerable. The data sources lack standardization and incorporate varying data formats, nomenclature, and organizational structure. As a result, data integration is a significant bottleneck to research productivity, anecdotally consuming up to 50% or more of the total data processing effort – effort that could be better spent analyzing data.

To demonstrate the utility of knowledge graphs, we first set about designing an ontology to model the contents of the core ATM data sources. Then we constructed a corresponding knowledge graph populated with instances that were derived from actual ATM data generated by these core sources during one month of air traffic operations in the New York metropolitan area. Finally, we worked with researchers to demonstrate how the knowledge graph could be queried (using SPARQL) to help answer active research questions. Some of the representative research-related queries we generated include the following:

- Find Newark Airport flight arrivals that passed through the PENNS airspace fix[1] and landed during rainy and windy conditions in July 2014;
- Find which airspace sector controlled the most flights in the US during the 9am Eastern hour on 7/15/14;

This paper is authored by an employee(s) of the United States Government and is in the public domain. Non-exclusive copying or redistribution is allowed, provided that the article citation is given and the authors and agency are clearly identified as its source.
WWW '19, May 13–17, 2019, San Francisco, USA
2019. IW3C2 ISBN 978-1-4503-6675-5/19/05
https://doi.org/10.1145/3308560.3317706

[1] An airspace navigation fix, or waypoint, is a named geospatial location defined at an intersection or point along a designated airway or above a surface landmark. Fixes are used in aircraft guidance and navigation.

- List all aircraft types that were used in commercial airline flights departing or arriving JFK Airport on 7/15/14;
- Find flights subject to ground delay advisories on 7/15/2014

This paper describes some of our experiences building an ATM knowledge graph to help answer these types of questions, and presents some of the challenges and pain points we faced. The next section begins by introducing the ontology and knowledge graph in more detail.

2 NASA's ATM Knowledge Graph

The NASA ATM Ontology (ATMONTO) defines key classes of entities pertaining to the US National Airspace System and the management of air traffic through that system. The primarily RDF-based [2] ontology describes a wide variety of aviation entities, and features more than 150 classes, 150 datatype properties, and 100 object properties. The ontology is fully documented in [1] and available for download with accompanying sample data [2].

Briefly summarizing, the ATMONTO classes represent:

- **airspace infrastructure entities**: airports, runways, terminals, airways, waypoints, air sectors, air traffic control facilities, and air control regions;
- **flight-related entities**: points of origin and destination, airline operators, flight plans, flight trajectories, aircraft, aircraft subsystems, and airframe manufacturers; and
- **flight operating conditions**: current and forecast airport weather conditions, systemwide ATM advisories, and routing constraints due to weather, facility, or other disruptions.

ATMONTO serves as the representational foundation for the NASA ATM Knowledge Graph (ATMGRAPH). ATMGRAPH is populated with over 38M instances and 260M triples derived from infrastructure, flight, and weather data collected for the three largest New York area airports (JFK, Newark, and LaGuardia) during July 2014. Included are data from approximately 100K flights arriving or departing from these airports during this month. The data are stored in an instance of OntoText's GraphDB triple store running at NASA.[3]

To give a flavor for how the ATM Knowledge Graph is structured, consider Figure 1. At the center of this graph fragment is a node representing a specific instance of a flight: UAL535 on 2014-07-15 departing at 00:19:00. The flight is linked to a variety of associated nodes via object properties: the specific aircraft flown (the aircraft with registration number N589UA, a Boeing 757 model 222), the carrier (United Airlines), the flight's departure and arrival airports (JFK and LAX), the planned and actual flight route between these airports (both represented as linked lists of either airways or actual trajectory points en route).

[2] ATMONTO uses a limited set of OWL constructs, including property restrictions.
[3] A subset of the data covering 100 actual flights has been released for documentary purposes [2]. The triple store is not currently accessible to the public.

The flight and the other instances in the graph fragment contain values for various datatype properties defined in ATMONTO. (Note: the datatype properties and values are not shown in the Figure.) For example, the flight has properties corresponding to the flight identifier ('UAL535'), the flight category ('commercial'), and the departure and arrival date/time. Although the knowledge graph depicted in Figure 1 focuses on the representation of a single flight, the actual graph contains data for 100K flights and is extremely densely connected. For example, over 38K flights are linked to JFK as either their departure or arrival airport during July 2015; and 59 flights flown during that period used the aircraft registered as N589UA. Furthermore, the fragment in Figure 1 illustrates only a small fraction of the total number of classes and link types found in the overall graph.

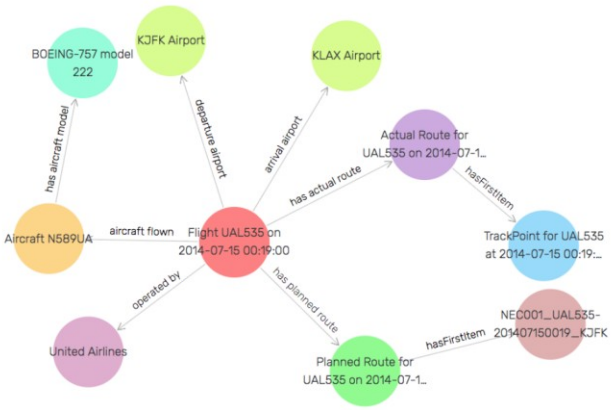


Figure 1: Fragment taken from ATMGRAPH illustrating some of the linkages centered around a flight. All nodes reside within one of ATMONTO's defined namespaces. Datatype properties -- such as flight start and end time -- are not shown.

ATMGRAPH is constructed principally from eight different structured data sources consisting of approximately 50 different low-level data products (e.g., database tables, custom data files, html files, spreadsheets). These data were produced by various governmental agencies (including the US Federal Aviation Administration (FAA), the US Department of Transportation, and the US National Oceanographic and Atmospheric Administration), non-governmental organizations (such as the International Civil Aviation Organization (ICAO)), and other web-based content providers (e.g. openflights.org). These structured sources are highly heterogeneous in format and structure. Python scripts and Java code were developed to transform these data from their original format into RDF triples conforming to the ATMONTO ontology, and the triples were loaded into the GraphDB triple store. As with all real-world applications, the data are noisy, contain errors, and must be extensively preprocessed to ensure the quality of the information in the knowledge graph.

In the next section we review some of the many practical challenges we faced in constructing ATMGRAPH.

3 Challenges

3.1 Entity Naming, Resolution, and Linking

Key components of the overall process necessary to transform the source data into linked data involve entity naming, resolution, and linking. Although these aspects are generally thought to be relatively straightforward when dealing with structured data sources versus unstructured text, they were non-trivial with the structured sources produced in the ATM domain. The difficulties relate back to the heterogeneity of data producers, data formats, and data encodings – and to the lack of standardization in data products overall. In the next subsections we give examples of these issues.

3.1.1 Assembling a Flight Instance . Although a flight would seem to be a key entity in ATM domain, there is no single authoritative source of flight information to be found within any of the core data sources used by the researchers or published by the FAA. Instead, information about flights must be pieced together by accumulating data from multiple sources in order to synthesize the properties of a flight instance and link it appropriately. A principal source of information we used to bootstrap the flight data amalgamation process was the ASDI (Aircraft Situation Display to Industry) flight track data, a source generated by FAA and used widely by commercial flight tracking web sites, such as flightaware.com and flightview.com. An ASDI file contains one line for each 60-second reporting period for every flight aloft during the timeframe covered by the file. The tracking data provides proof positive that a flight was flown, and provides some essential data such as the flight identifier, the departure/arrival airports, and tracking information (altitude, latitude, longitude, airspeed, etc.). But key additional information is missing from ASDI files, including the aircraft registration number, the aircraft manufacturer, and the airline carrier, among other items. Those data must be inferred from the encoded information and combined using auxiliary data dictionaries. For example, the airline carrier must be inferred from the flight identifier (e.g., in 'UAL535', 'UAL' is the encoding for United Airlines).

3.1.2 Naming the Flight Instance . The creation of a unique flight identifier is something that the aviation industry has struggled with for some time [3]. FAA and EUROCONTROL (FAA's counterpart in Europe) have come to an understanding over the past decade that integration of flight data from multiple sources requires the use of a Globally Unique Flight Identifier (GUFI), and are developing registry services for providing a GUFI. However, GUFIs are not yet in widespread usage within the aviation industry, creating challenges for entity resolution and making it difficult to determine which data from multiple data products pertains to the same flight.

3.1.3 Multiple Standards in Use . Even though many of the core ATM data products are generated by the same government agency (FAA), they do not consistently encode references to a given entity, and in some cases, there are multiple standards in use for the encoding [4]. As a simple example, the airport code for John F. Kennedy International Airport can be expressed using the ICAO standard (yielding 'KJFK'), the IATA (International Air Transport Association) standard ('JFK'), or the FAA standard, which in this case – but not always – is the same as the IATA code.

3.2 Spatial and Temporal Representation

Aviation information involves spatial and temporal aspects that must be adequately addressed by any representation [5]. Some examples of specific temporal requirements on entities in the ATM domain include: periodic updates to aircraft routes; time-limited air traffic initiatives; temporary airport obstacles; scheduled runway closures; and forecast weather phenomena. In addition to temporal aspects, many of the problems addressed by ATM researchers involve spatial reasoning. As a simple example, it is often necessary to determine whether an aircraft has passed through a defined region of the airspace, such as a flight control sector or a restricted airspace. Or it may be necessary to calculate the closest distance between an aircraft and a given navigation airway. Our representation of a flight trajectory (i.e., a flight path) illustrates some of the issues we faced, and the tradeoffs that must be made between expressivity and efficiency.

3.2.1 Representing a Flight Trajectory . A trajectory in ATMONTO is represented as a sequence of explicit track point instances. Each track point corresponds to a specific reporting time[4] when an aircraft's speed and navigation fix (its latitude, longitude, and altitude) are captured and relayed to ground systems. Unfortunately, this representation, while adequate for many needs, is verbose and leads to a proliferation of track points that undermines the efficiency of SPARQL query responses. Fully 70% of the 38M nodes in ATMGRAPH are either track points or navigation fixes, as described in [6]. Furthermore, performing the geometric calculations necessary to compute line crossings or geometric distance calculations in SPARQL is awkward and often impossible without escaping to custom-coded programming language functions.

An alternative and more efficient approach involves the use of a geospatial representation to capture the trajectory as a segmented line, employing something similar to the standards-based WKT (Well-Known Text) or GML (Geographic Markup Language) geometry representations used in geospatial databases, such as PostGIS and Oracle Spatial. These databases are optimized for performing fast spatial queries. This type of geometry representation is much less verbose than the representation used in ATMONTO; rather than being represented by separate instances in the knowledge graph, all of the track points in a trajectory would be specified using a single text string (e.g., '[$point1_{lat}$, $point1_{lon}$, $point1_{alt}$] [$point2_{lat}$, $point2_{lon}$, $point2_{alt}$] …') stored in a datatype property of a trajectory instance. We experimented with this type of representation as part of the GeoSPARQL facility supported by our triple store[5]. GeoSPARQL is the Open Geospatial Consortium's standard for representing and querying geospatial linked data [7]. But GeoSPARQL only represents 2-D geometries, and 3-D was necessary to represent the altitude dimension of the

[4] ATMONTO supports temporal aspects using either time points (modeled as datetime properties) or intervals (modeled as classes with start/end time points).
[5] GraphDB includes native GeoSPARQL support via a plugin.

flight trajectories and other three-dimensional regions of the airspace.

Even if GeoSPARQL supported 3-D geometries, there would still be an issue of representational granularity to consider. Because the compact geometry representation in GeoSPARQL buries the individual track point information within a text string stored in a datatype property, it is not possible to make statements about individual track points. For example, an important concept in ATM is the 'top of descent' – the spot at which an aircraft begins its descent to the airport. It would be preferable to mark a specific track point as the top of descent using a property on a track point instance. But with the GeoSPARQL representation, this is not possible. Similarly, if we want to record multiple aircraft passing through the same track point at different times or to record successive weather conditions at the track point, this would not be possible using GeoSPARQL.

3.3 Scaling

We have reported previously [6] about our efforts to benchmark performance of two different triple stores as we scaled up the amount of data in ATMGRAPH from one day to one month of New York flight operations. The results indicate that for 60% of our 17 benchmark queries, the execution time increases roughly linearly in the number of triples. (For one benchmark query, however, execution time increased exponentially.) In only 30% of the queries was execution time was not impacted by the increase in triples. Linear and exponential increases in execution time signal problems ahead as we attempt to scale up ATMGRAPH to include ever larger amounts of flight data. NASA aeronautics researchers sometimes perform multi-year ATM analyses across the entire US, so one month of data for New York is quite limiting. Hybrid approaches in which a portion of the flight data is stored in either a high-performance big data system or a geospatial database may help alleviate performance issues. Nevertheless, the performance of state-of-the-art triple stores is not sufficient to support real-time querying of the knowledge graph – at least for 70% of our benchmark queries. Furthermore, query optimization tools that might allow us to hand-tune performance of a SPARQL query engine are relatively primitive and difficult to use, at least in our experience.

3.4 Visualization

Visualization plays an important exploratory role for those knowledge graph users who wish to closely examine query results and navigate through specific chains of instances in the graph. Our ATM researchers often wish to examine anomalous flights that are being flown outside of normal operating parameters, or flights flown under specific operating conditions, such as severe convective weather. In these circumstances, it would be useful for users to examine specific flight instances in ATMGRAPH as an adjunct to other analyses they are performing. Our experience with current tools is that visualization remains challenging in large, densely-connected graphs, a fact that has been well-documented [8–10]. Navigating through a graph in which hundreds or thousands of links emanate from a single node requires specific filtering techniques that allow users to selectively choose the link or links they wish to traverse and the nodes they wish to view. While the GraphDB environment we are using implements some filtering capabilities (and even allows users to write their own filters using SPARQL) these are not sufficient to support our users' graph exploration needs.

4 Related Work

There is a body of published work on the use of ontologies, and more recently knowledge graphs, in aviation-related applications, including ATM [11-14], aviation data management [15], aviation safety [16, 17], and avionics [18]. The Graph of Things [19] incorporates flight information and track points similar to the data incorporated in ATMGRAPH. Despite this related work, reuse of other ontologies was not generally practical in our real-world application and setting. In some cases, ontology details were not published; in other cases, ontologies were too simplistic and/or mismatched to the needs of our ATM application in terms of scope or level of detail. And finally, often the effort involved in locating, augmenting, and reusing ontologies is significant and outweighs their overall utility.

5 Summary

This paper has described our efforts toward developing a knowledge graph resource for the air traffic management community, including some of the challenges we faced in producing a prototype system for use by aerospace researchers. Of the challenges described, scaling is the most serious barrier to deployment, as a narrowly-scoped knowledge graph will be of limited use to NASA researchers. Further, without more intuitive knowledge graph query languages and visualization tools, non-experts will struggle to use the technology to its full potential.

6 ACKNOWLEDGMENTS

This work was funded by the Airspace Operations and Safety Program from within NASA's Aeronautics Research Mission Directorate. My thanks to the members of the Sherlock Data Warehouse Team at NASA Ames, who supported me in conducting this work, and the aerospace researchers who kindly provided their input. I also want to thank the FAA and the many other aviation data providers who made their data available for this effort.

REFERENCES

[1] Keller, R.M. *The NASA Air Traffic Management Ontology: Technical Documentation.* Technical Memo NASA/TM-2017-219526, National Aeronautics and Space Administration, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/2017 0006095.pdf, 2017.

[2] Keller, R.M. *ATMONTO Ontology.* https://data.nasa.gov/ontologies/atmonto/, release dated March 2018.

[3] EUROCONTROL. *Globally Unique Flight Identifier (GUFI) Requirements*,2014, https://www.fixm.aero/documents/GUFI%20Requirements%20v2 %201_Final.pdf.

[4] *Airport Codes.* https://www.aircharteradvisors.com/airport-codes/ Accessed 2/19.
https://fixm.aero/documents/GUFI%20Requirements%20v2%201_Final.pdf, 2014.

[5] EUROCONTROL. *AIXM 5 Temporality Model.* http://aixm.aero/sites/aixm.aero/files/imce/AIXM51/aixm_temporality_1.0.pdf , 2007.

[6] Keller, R.M., Ranjan, S., Wei, M.Y. and Eshow, M.M. *Semantic Representation and Scale-up of Integrated Air Traffic Management Data.* ACM, in Proceedings of the International Workshop on Semantic Big Data, San Francisco, 2016.

[7] Open Geospatial Consortium. *OGC GeoSPARQL - A Geographic Query Language for RDF Data.* Document # OGC 11-052r4, 2012.

[8] Gómez-Romero, J., Molina-Solana, M., Oehmichen, A. and Guo, Y. Visualizing large knowledge graphs: A performance analysis. *Future Generation Computer Systems*, 89 (2018), 224-238.

[9] Pienta, R., Abello, J., Kahng, M. and Chau, D. H. *Scalable graph exploration and visualization: Sensemaking challenges and opportunities.* IEEE, in Big Data and Smart Computing (BigComp), 2015.

[10] Wills, G. J. NicheWorks—interactive visualization of very large graphs. *Journal of computational and Graphical Statistics*, 8, 2 (1999), 190-212.

[11] Burgstaller, F., Steiner, D., Schrefl, M., Gringinger, E., Wilson, S. and van der Stricht, S. *AIRM-based, fine-grained semantic filtering of notices to airmen.* In IEEE *Integrated Communication, Navigation, and Surveillance Conference (ICNS)*,2015, pp. D3-1-D3-13.

[12] Neumayr, B., Gringinger, E., Schuetz, C. G., Schrefl, M., Wilson, S. and Vennesland, A. *Semantic data containers for realizing the full potential of system wide information management.* In IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), pp. 1-10. IEEE, 2017.

[13] van Putten, B.-J., Wolfe, S. R. and Dignum, V. An Ontology for Traffic Flow Management. In Proc. 8th Aviation Technology, Integration, and Operations Conference, AIAA, 2008.

[14] Vennesland, A., Neumayr, B., Schuetz, C. G. and Savulov, A. *Experimental ontology modules formalizing concept definition of ATM data.* SESAR Exploratory Research Report, http://www.project-best.eu/downloads/D1.1%20Experimental%20ontology%20modules%20formalising%20concept%20definition%20of%20ATM%20data.pdf, 2017.

[15] Keller, R. M. *Ontologies for Aviation Data Management.* In Proc. IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016.

[16] Ledvinka, M., Lališ, A. and Křemen, P. Toward Data-Driven Safety: An Ontology-Based Information System. *Journal of Aerospace Information Systems*, 16, 1 (2018), 22-36.

[17] Zhao, Q., Li, Q. and Wen, J. Construction and application research of knowledge graph in aviation risk field. In Asia Conference on Mechanical and Aerospace Engineering, EDP Sciences, 2017.

[18] Blasch, E. Ontologies for nextgen avionics systems. In Proc. IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), 2015.

[19] Le-Phuoc, D., Quoc, H. N. M., Quoc, H. N., Nhat, T. T. and Hauswirth, M. The Graph of Things: A step towards the Live Knowledge Graph of connected things. Journal of Web Semantics, 37 (2016), 25-35.

Scalable Knowledge Graph Construction over Text using Deep Learning based Predicate Mapping

Aman Mehta*
International Institute of Information Technology Hyderabad
India
aman.mehta@research.iiit.ac.in

Aashay Singhal*
International Institute of Information Technology Hyderabad
India
aashay.singhal@research.iiit.ac.in

Kamalakar Karlapalem
International Institute of Information Technology Hyderabad
India
kamal@iiit.ac.in

ABSTRACT

Automatic extraction of information from text and its transformation into a structured format is an important goal in both Semantic Web Research and computational linguistics. Knowledge Graphs (KG) serve as an intuitive way to provide structure to unstructured text. A fact in a KG is expressed in the form of a triple which captures entities and their interrelationships (predicates). Multiple triples extracted from text can be semantically identical but they may have a vocabulary gap which could lead to an explosion in the number of redundant triples. Hence, to get rid of this vocabulary gap, there is a need to map triples to a homogeneous namespace. In this work, we present an end-to-end KG construction system, which identifies and extracts entities and relationships from text and maps them to the homogenous DBpedia namespace. For Predicate Mapping, we propose a Deep Learning architecture to model semantic similarity. This mapping step is computation heavy, owing to the large number of triples in DBpedia. We identify and prune unnecessary comparisons to make this step scalable. Our experiments show that the proposed approach is able to construct a richer KG at a significantly lower computation cost with respect to previous work.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; **Neural networks**; • **Information systems** → *Entity resolution*; *Ontologies*.

KEYWORDS

Knowledge Graph, Predicate Mapping, Deep Learning, Scalability, Sentence Simplification

ACM Reference Format:
Aman Mehta, Aashay Singhal, and Kamalakar Karlapalem. 2019. Scalable Knowledge Graph Construction over Text using Deep Learning based Predicate Mapping. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3317708

*represents equal contribution.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317708

(a)	*Barack Obama was born in Honolulu.*
(b)	*Obama was elected in 2009 as the president of the United states. He belongs to Honolulu.*
(c)	*Barack served as the 44th president of the United States and grew up in Honolulu.*

Table 1: Example text (referred in multiple sections)

1 INTRODUCTION

Motivation. In any industry, text documents that contain important data are a common occurrence. Hence, improving the ability of machines to understand the intent and context of information to the level of humans is one of biggest challenges faced today. Achieving this would involve transformation of unstructured text to a structured format. Knowledge graphs provide an intuitive way of giving shape and structure to the initially unstructured information. A fact in a KG is represented by a triple of the form <*S;P;O*>. *S*, *O* denote the *subject* and *object*, respectively and *P* is the *predicate* which describes the relationship between *S* and *O*.

A KG built over any unstructured text document helps in making the information in the document queryable. This makes the previously inoperable information usable for tasks such as search result ranking, recommendation, question answering, etc [2, 13, 30, 31, 37]. Another interesting application of constructing KG over text is to add new or missing information to an existing KG like DBpedia [2], Freebase [4].

Traditional methods ([9]) focus on building KGs from infobox templates and categorization information in the Wikipedia articles. However, the unstructured text of these articles is left unprocessed. In order to make this natural language text structured and usable, generating a KG over text becomes an important task. Previous works such as [3, 6, 7, 11, 20, 36] use information extraction system to extract facts from NL text. But these facts do not necessarily follow the paradigm of a KG (such as DBpedia), making the KG construction task challenging. For instance, consider the output of sample sentence *(a)* from (table 1) using OLLIE information extractor - (Barack Obama, was born in, Honolulu). This fact is identical to (Barack Obama, birthPlace, Honolulu) in DBpedia. Although these two facts are identical, there is a vocabulary gap between them. In the fact extracted from text, *"was born in"* is a natural language phrase, whereas *"birthPlace"* is a formatted predicate in DBpedia. These two relational phrases have different surface forms and hence they cannot be mapped just on the basis of string similarity.

Furthermore, same DBpedia entities and predicate could be used in multiple NL excerpts. For instance, *"He"*, *"Barack Obama"* and

"Barack" in the facts extracted for *(a), (b)* and *(c)* (from table 1) correspond to *"Barack_Obama"* in DBpedia. Similarly, the predicates such as *"was born in"*, *"grew up in"*, *"belongs to"* are related to the same predicate *"birthPlace"* in DBpedia. As shown in fig 1, this leads to a high amount of redundancy in the constructed KG, giving rise to an unnecessarily large sized KG. This redundancy leads to wastage of space for its storage and wastage of time in execution of graph retrieval algorithms [22, 23, 38]. Therefore, it is important to efficiently resolve entities and their relationships in triples to facilitate a queryable Knowledge Graph. The most common approach to resolve them is by mapping them to a single homogeneous namespace, such as DBpedia namespace.

A number of studies have proposed ideas for Entity Mapping [25]. However, only a few studies have worked on the task of Predicate Mapping. A few rule-based [10] and similarity-based approaches [19] have been proposed in recent years, but both of them have their own limitations. Simple rule-based approaches cannot generate rules efficiently, especially when text sources are sparse. This is due to the fact that these rules are manually generated and hence it cannot get rid of all redundancies. On the other hand, similarity-based solutions for Predicate mapping are challenging in two aspects:
a) To map a predicate to another namespace, they need to capture the accurate semantics for calculating similarity scores
b) comparing a predicate to each of the predicate in DBpedia is highly time consuming owing to the number of *candidates* (i.e. the number predicates in DBPedia namespace to which a particular text predicate can map to). T2KG [19] lacks in both of these aspects. Firstly, it uses a word2vec based model which is not able to capture all semantic features since it uses shallow neural networks[1]. Secondly, it scans all DBpedia triples in order to map a single predicate of KG to DBpedia namespace[2]. So, when the text sources are large, it becomes a time consuming task to map all text predicates to DBpedia predicates. Even tougher when the size of the DBpedia itself is as large as 1.3 billion[3] triples, such as in the case of English version of DBpedia. We address both of the above challenges in the proposed approach.

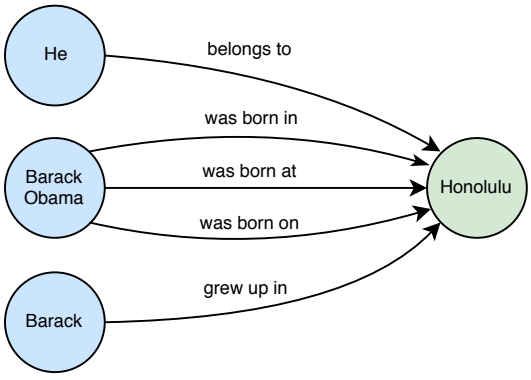


Figure 1: Knowledge Graph obtained using information extraction (OLLIE [36])

[1]https://en.wikipedia.org/wiki/Word2vec
[2] [19] Page 5: eq. (2) and eq. (3)
[3]https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10

Contributions. In this work, we introduce an end-to-end KG construction system from unstructured text. In order to make this KG more useful, we map our KG to DBpedia namespace far more efficiently than the previous approaches. Under the Predicate Mapping step, we implement a pruning strategy to reduce a large number of unnecessary comparisons in order to make this step scalable. The end-to-end system is able to build a larger KG, with less redundancy while significantly reducing the computation cost in finding the Predicate Mapping step. Our main contributions are as follows:

- We introduce a system to construct KG from text which is consistent with the DBpedia namespace.
- We propose a Deep Learning model to calculate similarity between two predicates for the Predicate Mapping step.
- We develop optimization strategy to reduce the *Candidate Set* (described in section 3.2) for each text predicate.
- We introduce a sentence simplification component to improve triple extraction.
- We perform multiple experiments to evaluate our KG construction system with other systems. The experiments show the effectiveness of our Predicate Mapping component and the benefits of the Sentence Simplification component.
- A separate study to quantify the extent of redundancy reduction in different domains of Wikipedia articles in the KGs constructed using our system.

The rest of this paper is structured as follows: section 2 describes different components of our pipeline, section 3 describes in detail our Predicate Mapping model and section 4 describes the training details of our model. We perform experiments and show our results in section 5 and conclude our work in section 6.

2 END-TO-END PIPELINE: TEXT TO KG

In this section, we describe in detail, the steps to construct a Knowledge Graph from unstructured text. Our end-to-end pipeline has six components: The *Entity Mapping* component maps the entities in the text triples to the DBpedia entities. The *Sentence Simplifier* which simplifies sentences before triple extraction step to overcome limitations of triple extraction on complex sentences. The *Co-reference Resolution* component which finds and replaces all the expressions that refer to the same entity in the text. The *Triple Extractor* that uses information extraction techniques to extract relation triples from text (called text triples). The *Metadata Processing* component prepares and stores the set of candidate DBpedia predicates for a given text predicate, that could possibly be the mapping. The *Predicate Mapping* component maps a text triple's predicate to its matching predicate in DBpedia namespace.

Entity Mapping: An important step in KG construction is linking named entities to unique identifiers. In this component, we use a named entity recogniser [25] to mark all the named entities in the text. Now, for an entity which has a possible mapping in DBpedia, we use the URI of such a DBpedia entity as the representation of the text entity. Otherwise, we define a custom namespace to create their URIs. The output from Entity Mapping step are the same set of sentences, but entities replaced by the URIs as defined by above conditions. For eg: *"Barack Obama"*, *"Obama"* and *"Barack"* will be replaced by *DBpedia:Barack_Obama*, which is the URI of Barack

Obama entity in DBpedia. Hence, output of this layer for above three instances will be,
(a) DBpedia:Barack_Obama was born in DBpedia:Honolulu.
(b) DBpedia:Barack_Obama was elected in 2009 as the president of the DBpedia:United_States. He belongs to DBpedia:Honolulu.
(c) DBpedia:Barack_Obama served as the 44th president of the DBpedia:United_States and grew up in DBpedia:Honolulu.

Sentence Simplification: English is a difficult language for machines to comprehend. As and when sentences grow complex in nature, extraction of relation triples from unstructured text becomes tougher. Due to complex structure of sentences, the frameworks which try to convert unstructured text into a relation triple, cease to identify correct relationships between the entities at hand. Hence, the aim of this component is to identify complex sentences and convert them into simpler sentence(s), such that the semantics is consistent even after simplification. We implement the approach proposed in [28] to achieve this. Here, *(a)* (table 1) remains unchanged, because it is not a complex English sentence whereas *(b)* and *(c)* change to:
(b) DBpedia:Barack_Obama was elected in 2009. DBpedia: Barack_Obama was the president of the DBpedia:United_States. He belongs to DBpedia:Honolulu.
(c) DBpedia:Barack_Obama served as the 44th president of the DBpedia:United_States. DBpedia:Barack_Obama grew up in DBpedia:Honolulu.
In section 5.3, we perform a separate study to evaluate the impact of this component in our system.

Co-reference Resolution: There are a lot of expressions for a given entity in unstructured text, such as pronouns and abbreviations, which act as a proxy for some real-world entity. And if these expressions are left unannotated, we may lose crucial information. Hence it is important to group all the mentions of an entity and link them to an URI. We use approach proposed by [34] to determine and replace all the linguistic expressions which refer to the same real-world entity and link them to its URI, as determined by the Entity Mapping component. The output of this layer replaces all the co-reference chains by their URI. Only *(b)* changes, since it is the only sentence with a co-reference chain.
(b) DBpedia:Barack_Obama was elected in 2009. DBpedia:Barack_Obama was the president of the DBpedia:United_States. DBpedia:Barack_Obama belongs to DBpedia:Honolulu.

Triple Extraction: In this component, we extract relation triple from plain text. This is the primary step in extracting information from unstructured text. A relation triple is a data entity composed of a subject-predicate-object. Subject and object are real-world entities and predicate describes the relationship that the subject entity has with object entity. Below we show the triples extracted from the portion of text that talk about the *birth place* of *Obama*.
(a) {<DBpedia:Barack_Obama; was born in; DBpedia:Honolulu>, <DB-pedia:Barack_Obama; was born at; DBpedia:Honolulu>, <DBpedia:Barack_Obama; was born on; DBpedia:Honolulu>}
(b) {<DBpedia:Barack_Obama; belongs to; DBpedia:Honolulu>}
(c) {<DBpedia:Barack_Obama; grew up in; DBpedia:Honolulu>}
The task of extracting a relation triple from plain text can be carried out by any information extraction [1, 8, 12, 24, 39] technique. These techniques extract relation triples from text by identifying relation phrases and associated arguments in a sentence without requiring a pre-specified vocabulary. We use Open Language Learning for Information Extraction (OLLIE) [36] as the triple extractor in our implementation.

Metadata Processing: This step is a preprocessing step for our main task of Predicate Mapping (discussed in section 3). The aim of this component is to store all possible DBpedia predicates to which a predicate from text triple can map, by using a pruning strategy. To find such a set of possible DBpedia predicates for a given text predicate, later defined as *Candidate Set*, we define two sub-tasks. Let our text triple be of the form R_t =<S;P_t;O>. At first, we need to find the class to which S and O entities belong. If the S and O are mapped to some entities in DBpedia, we extract S_class and O_class using DBpedia class information. Otherwise, we use a Named Entity Recognizer (NER) [14] to predict it's class and map it to a DBpedia class using NER and Disambiguation (NERD) ontology [35]. For example, an NER class *Person* maps to its equivalent DBpedia class, *DBpedia:Person*. Secondly, after the subject and object class of a text triple are extracted correctly, we prune the DBpedia predicates which can surely not be the mapping for P_t. The idea is, if the relation triple in the text R_t, is defined between a *Person* class entity and a *Place* class entity, clearly the map of P_t in DBpedia namespace will also be only among the predicates which have domain as *DBpedia:Person* and range as *DBpedia:Place*. So, instead of evaluating its similarity score with all the predicates of DBpedia, we only do it for relationships between a *DBpedia:Person* and a *DBpedia:Place* class. This remaining set, after pruning the unimportant predicates, is defined as the *Candidate Set (CS)* (described in section 3.2) for P_t. We explain in detail (section 3.3) the difference in cost of computation of previous similarity-based approach and our approach.

Predicate Mapping: The aim of this component is to map predicate of a text triple to its matching predicate in DBpedia. From the previous component we obtained the *Candidate Set (CS)* for P_t. Here, we find the similarity of the text predicate with every candidate in the *CS* using the model explained in next section. We finally map P_t to the most similar candidate predicate in DBpedia. Our model is able to map *"was born in", "belongs to", "was born at", "was born on"* and *"grew up in"* correctly to *DBpedia:birthPlace*. Hence, the output for the examples becomes:
(a) {<DBpedia:Barack_Obama; DBpedia:birthPlace; DBpedia:Honolulu>}
(b) {<DBpedia:Barack_Obama; DBpedia:birthPlace; DBpedia:Honolulu>}
(c) {<DBpedia:Barack_Obama; DBpedia:birthPlace; DBpedia:Honolulu>}
Simply extracting relation triples using OLLIE (as shown in fig 1) will give us a graph with four nodes (entities: {*"Barack Obama", "He", "Barack", "Honolulu"*}) and five edges (predicates: {*"was born at", "was born in", "was born on", "belongs to"*}) which clearly holds redundant information. Our approach is able to resolve all the identical mentions of entities and predicates hence forming a more precise graph with two nodes and one edge between them, i.e. *<DBpedia: Barack_Obama;DBpedia: birthPlace;DBpedia: Honolulu>*. The output of the Predicate Mapping step for a particular text triple is a triple that has its entities and predicate linked to DBpedia. The collection of such triples is our constructed KG. Thus, our system has constructed a KG over a specific text article or document. This KG makes the raw text information available for document specific tasks such as question answering and information retrieval. Furthermore, if there is some information present in this text, but not

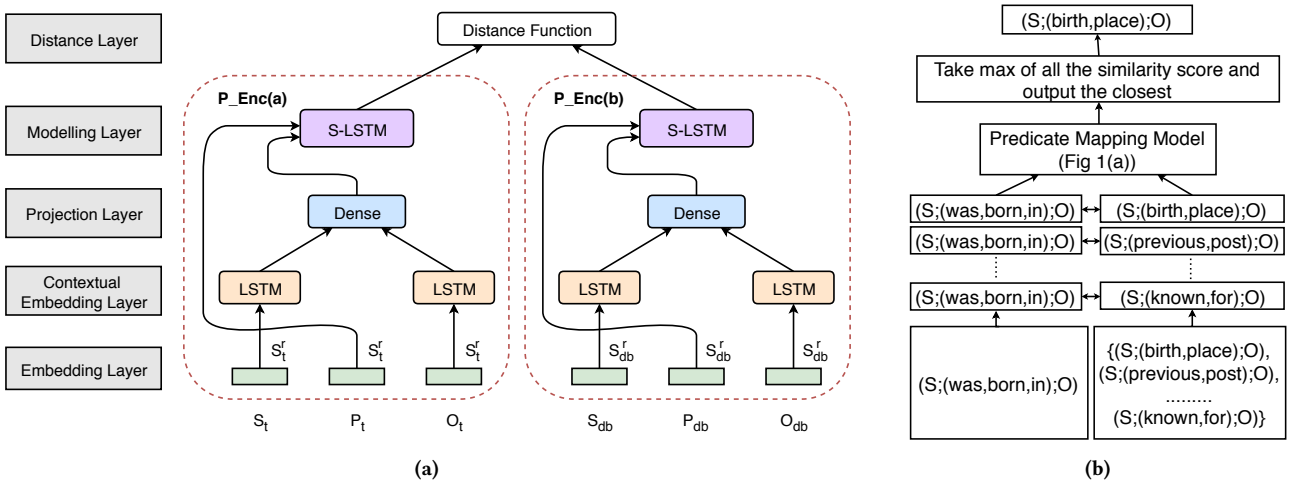


Figure 2: (a) End-to-end Predicate Mapping model, (b) An illustrative example of Predicate Mapping, where S and O are Subject Class and Object Class respectively

in DBpedia then we could add the missing information to DBpedia, by merging it with our constructed KG.

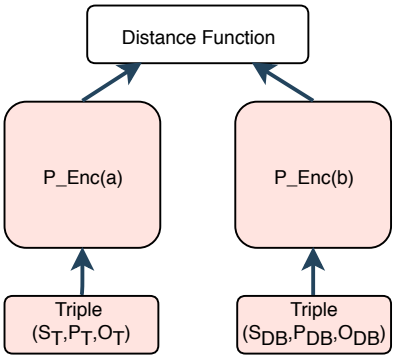


Figure 3: Overview of Predicate Mapping Model

3 PREDICATE MAPPING

Let $<S_{db}; P_t; O_{db}>$ denote a relation triple extracted from the text using the above pipeline. Our aim is to find a DBpedia predicate (say P_{db}) which is semantically similar to P_t and map P_t to P_{db}. We use a deep learning based similarity approach to find this P_{db}.

The idea is to get a vector representation of the text predicate and compare its similarity with possible DBpedia candidates. The representation of the predicate is formed using the Predicate Encoder ($P_Enc(a)$ and $P_Enc(b)$ as shown in fig. 2). The predicate encoder is a multi-stage architecture (described in Section 3.1) which uses subject and object class as initial context to generate the representation since they store valuable information about the predicate.

Let the final encoded vector representation of the two predicates obtained from the predicate encoder be $H_t = [x_1, .., x_k]$ and $H_{db} = [y_1, .., y_k]$. The similarity between the two vector representations H_t and H_{db} is given by the following equation:

$$f(H_t, H_{db}) = exp(-||H_t - H_{db}||) \in [0, 1] \quad (1)$$

In the next section, we define our Predicate Mapping model.

3.1 Model

Our Predicate Mapping model (shown in fig. 2(a)) is a hierarchical multi-stage Siamese Network, inspired by [27]. The model consists of five layers. The *Embedding Layer* maps each word to a vector space using a pre-trained embedding model. The *Contextual Embedding Layer* utilizes the vector representations of surrounding words to refine the embeddings of the words. The *Projection Layer* forms the initial state for our Modeling Layer by densely combining subject class and object class embeddings. The *Modeling Layer* employs a Siamese Recurrent Network to create final embedding of the predicate using the combined embedding of subject class and object class as context. The *Distance Layer* computes the final similarity score.

Embedding Layer: This is a word embedding layer which maps each word to a high-dimensional vector space. In our model, we use d-dimensional pre-trained GloVe word embeddings [32]. Let $\{s_1, s_2, ...s_{k_t}\}$, $\{p_1, p_2, ...p_{l_t}\}$ and $\{o_1, o_2, ...o_{m_t}\}$ denote the sequence of words in subject, predicate and object of a text triple and similarly for a DBpedia triple. The output of this layer consists of three matrices for text triple: $S_t^r \in \mathbb{R}^{d \times k_t}$, $P_t^r \in \mathbb{R}^{d \times l_t}$, $O_t^r \in \mathbb{R}^{d \times m_t}$.

$$S_t^r = [s_1^r s_2^r ... s_{k_t}^r]$$
$$P_t^r = [p_1^r p_2^r ... p_{l_t}^r]$$
$$O_t^r = [o_1^r o_2^r ... o_{m_t}^r]$$

where $s_i^r \in \mathbb{R}^{d \times 1}$, $p_i^r \in \mathbb{R}^{d \times 1}$ and $o_i^r \in \mathbb{R}^{d \times 1}$ represent the embedding of s_i, p_i and o_i. Similarly for DBpedia triple, we obtain: S_{db}^r, P_{db}^r and O_{db}^r.

Contextual Embedding Layer: In this layer, we use LSTM [17] to output the encoded representation of a given sequence. We use four LSTMs, two for S_t and S_{db} to form subject class encodings and the other two for O_t and O_{db} to form object class encodings.

LSTMs adapt feedforward neural networks for sequence data where at each time-step (say x), updates to a hidden state vector h_x are performed. In our case, the input sequence data is, $[s_1^r s_2^r ... s_{k_t}^r]$. These updates also rely on a memory cell containing four components (which are real-valued vectors): a memory state c_x, an output gate q_x (that determines how the memory state affects other units), an input gate i_x (that controls what gets stored in memory) and a forget gate f_x (that controls what does not get stored in memory). Below are the updates performed (for S_t) at each time-step $x \in 1, ..., k_t$:

$$\begin{aligned} i_x &= sigmoid(W_i s_x^r + U_i h_{x-1} + b_i) \\ f_x &= sigmoid(W_f s_x^r + U_f h_{x-1} + b_f) \\ \widetilde{c_x} &= tanh(W_c s_x^r + U_c h_{x-1} + b_c) \\ c_x &= i_x \cdot \widetilde{c_x} + f_x \cdot c_{x-1} \\ q_x &= sigmoid(W_q s_x^r + U_q h_{x-1} + b_q) \\ h_x &= q_x \cdot tanh(c_x) \end{aligned} \quad (2)$$

where $W_i, W_f, W_c, W_q, U_i, U_f, U_c, U_q$ are the trainable weight matrices and b_i, b_f, b_c, b_q are the trainable bias-vectors. We take the last hidden state of the LSTM (i.e. h_{k_t}) as our final representation. Hence, we obtain $G_t \in \mathbb{R}^{d\times 1}$, $U_t \in \mathbb{R}^{d\times 1}$ for subject classs (S_t) and object class (O_t) of text triple and $G_{db} \in \mathbb{R}^{d\times 1}$, $U_{db} \in \mathbb{R}^{d\times 1}$ for subject class (S_{db}) and object class (O_{db}) of DBpedia triple.

Projection Layer: This layer aims to prepare context for Modeling Layer using G and U vectors. We define

$$J_t = \{G_t; U_t\}$$
$$J_{db} = \{G_{db}; U_{db}\}$$

as the concatenation of subject class and object class representation of text triple and DBpedia triple respectively. This layer maps $J_t \in \mathbb{R}^{2d\times 1}$ to $Z_t \in \mathbb{R}^{d\times 1}$ by applying a Dense Layer and similarly $J_{db} \in \mathbb{R}^{2d\times 1}$ to $Z_{db} \in \mathbb{R}^{d\times 1}$. Z_t and Z_{db} store the combined representation of subject-object class of text and DBpedia triple and serve as the context for finding the representation of their respective predicates.

Modeling Layer: In this layer, we use another pair of LSTMs to create encoded representation of text predicate and candidate predicate. We use the output of the previous layer as context, i.e., Z_t and Z_{db} become the first hidden states (h_0 in eq. (2)) for the two LSTMs respectively. The input to this LSTM is P_t. We obtain $H_t \in \mathbb{R}^{d\times 1}$ as the last hidden state of this *S-LSTM* (as shown in fig. 2(a)). Similary, we obtain $H_{db} \in \mathbb{R}^{d\times 1}$. H_t and H_{db} represent the final text and DBpedia triple encoding, respectively.

Distance Layer: This layer computes the similarity score between H_t and H_{db} using the eq. (1). This score is used to choose the most similar predicate amongst a set of predicates called *Candidate Set* (discussed in detail in the next section).

3.2 Candidate Selection

In the previous work, the approach to map predicate of a text triple to DBpedia predicate requires high computation cost. The high computation is because it requires an entire scan of DBpedia[4]. Thus making the computation cost in the order of total number of triples in DBpedia (1.3×10^9). This step has many unnecessary comparisons, for example, an extracted predicate "*was born in*" is being compared with DBpedia predicates such as *DBpedia:biggestCity* and *DBpedia:altitude*. To overcome this, we identify the DBpedia predicates which can be the possible mapping for the text triple's predicate.

The idea is as follows: each predicate, say P_{db} in DBpedia has a unique subject class and object class, defined in the RDF-schema as *rdfs:domain* and *rdfs:range* respectively. This defines that, in any relation triple which has P_{db} as it's predicate, say <$S; P_{db}; O$>, all such S's and O's belong to the class defined by *rdfs:domain* and *rdfs:range*, respectively. We exploit this information about a text triple, to narrow down our search space significantly. DBpedia ontology forms a subsumption hierarchy [21], hence our new search space does not lose any potential candidates. Thus, we introduce a notion of *Candidate Set (CS)* for a text triple.

Formally, a *CS* for a text triple, say R_t =<$S; P_t; O$>, is a set of all the DBpedia predicates (stored as <$S_class; P_{db}; O_class$> in the *CS*) whose predicates occur between S's class and O's class. In our experiments, we find that the average size of the *Candidate Set* for a predicate is approximately 200, which reduces the search space by an order of 10^6 for each text predicate. Now that we have defined *CS* for each extracted predicate, we describe the steps to output the *closest* (most similar) DBpedia predicate for a particular R_t. Consider C_t =<$S_class; P_t; O_class$> and the *Candidate Set* for P_t,

$$\begin{aligned} CS = \{&<S_class; P_{db}^1; O_class>, \\ &...., <S_class; P_{db}^K; O_class>\} \end{aligned} \quad (3)$$

where K is the size of the *Candidate Set*.

Let us do a running example for *Candidate Selection*. Consider a R_t and its corresponding C_t and *Candidate Set (CS)* be

R_t = <*DBpedia:Barack_Obama; was born in; DBpedia:Honolulu*>
C_t = <*DBpedia:Person; was born in; DBpedia:Place*>
CS = {<*DBpedia:Person; DBpedia: birthPlace; DBpedia:Place*>,
<*DBpedia:Person; DBpedia: previousPost; DBpedia:Place*>,
... ,<*DBpedia:Person; DBpedia:knownFor; DBpedia:Place*>}

where K = 255 since there are 255 different predicates between *DBpedia:Person* and *DBpedia:Place*[5]. The steps are:

1. We calculate the similarity score for C_t and i^{th} element of the *Candidate Set* ($CS[i]$), $\forall i \in [1, K]$ by feeding C_t and $CS[i]$ to our Predicate Mapping model. That is, we calculate similarity score of <*DBpedia:Person; was born in; DBpedia:Place*> with <*DBpedia:Person; DBpedia:birthPlace; DBpedia:Place*> and so on with all the elements of *CS*.
2. We find the candidate from the *CS* which gives maximum similarity score with C_t, say $CS[j]$. In our example, we find that <*DBpedia:Person; DBpedia:birthPlace; DBpedia:Place*> has the maximum similarity with <*DBpedia:Person; was born in; DBpedia:Place*>.
3. Finally, we map P_t with the predicate of $CS[j]$ by replacing all occurrences of P_t with P_{db}^j. That is, we update our R_t to <*DBpedia:Barack_Obama; DBpedia:birthPlace; DBpedia:Honolulu*>.

[4] [19] Page 5: eq. (2) and eq. (3)

[5] We use SPARQL endpoint (http://yasgui.org/) to query DBpedia

So effectively, we compare *"was born in"* with only 255 predicates in DBpedia. We input each pair C_t and $CS[i]$ to our model (as shown in fig. 2(b)) and get similarity scores for each of them. The vector representation of *"was born in"* using *Person* and *Place* as context comes out to be closest (using eq. (1)) to the vector representation *DBpedia:birthPlace* calculated using *Person* and *Place* as context. If the size of *CS* for any text predicate is zero, we discard that text triple.

3.3 Scalability

Consider a text triple R_t =<S_t; P_t; O_t> and let its mapping be R_{db} =<S_{db}; P_{db}; O_{db}>. In the previous similarity based approach [10], it does an entire scan of DBpedia to find[4] this P_{db}. We find that it makes a lot of unnecessary computations. In our approach, we get rid of them by defining our *Candidate Set* using SPARQL queries. *As per the latest release of DBpedia*[6] *, in order to map a single P_t to P_{db}, our system requires approximately 200 computations while T2KG[19] essentially requires computation in the order of size of DBpedia, that is,* 1.3×10^9.

4 TRAINING DETAILS

To train this model in an end-to-end fashion, one would require a dataset consisting of pairs of similar triples. But, to the best of our knowledge, there is no such dataset available. Therefore, we need to train the components independently. We train two different models on two different datasets: SNLI[5] corpus and SemEval STS [18]. The SNLI dataset is a set of human-written English sentence pairs manually labeled for the task of Recognizing Textual Entailment (RTE). While the SemEval STS dataset consists of labeled cross-level semantically similar pairs of a paragraph and a sentence. The Predicate Mapping model is divided into three parts for training.

4.1 Contextual Embedding Layer Training

This part aims at obtaining the weights ($W_i, W_f, W_c, W_q, U_i, U_f, U_c$, U_q in eq. (2)) for the Contextual Embedding Layer. We train the LSTM in this layer similar to a Language Model (LM)[26]. The goal of an LM is to predict the $(k+1)^{th}$ word of sequence, given the previous k words. A huge chunk of text is given as input to the LSTM to train it as an LM. This way, it learns to encode a sequence of words and hence we use this trained LSTM as a sequence encoder in the Contextual Embedding Layer. We transform the sentences in the dataset into a sequence of tokens or words that is used for training the model. The weights obtained by training this model are used in all the four LSTMs in the Contextual Embedding Layer.

4.2 Projection Layer Training

This part aims at obtaining the weights for the Projection Layer. Since this layer, at its core, aims to reduce the dimension of the output of the Contextual Embedding Layer from $2d$x1 to dx1, we use an autoencoder[16] for this task. The input to the autoencoder is the concatenation of the two dx1 outputs of the LSTMs of the previous layer, which gives us a $2d$x1 vector. We store such $2d$x1 vectors for multiple triples and use them to train the autoencoder. The input is first transformed into a dx1 vector using a weight matrix $W_1 \in \mathbb{R}^{2d \times d}$, and then this intermediate vector is converted back to a $2d$x1 vector using another weight matrix $W_2 \in \mathbb{R}^{d \times 2d}$. The first part of the autoencoder reduces the input to dx1 vector. Hence, we use this W_1 matrix, as our weight matrix for this layer.

4.3 Modeling and Distance Layer Training:

This part aims at obtaining the weights of the Modeling Layer. We train the LSTMs in Modeling Layer and the Distance Layer as an end-to-end Siamese Network[29]. For a given pair of sentences in the dataset, first and second sentence are respectively fed to the two *S-LSTM*s. In the Distance Layer, similarity score between the last hidden states of the two *S-LSTM*s (H_t and H_{db}) is calculated using eq. (1). With similarity label known, the corresponding contrastive loss[15] is back propagated to the *S-LSTM*s according to the equation:

$$L(Y, H_t, H_{db}) = (Y)\frac{1}{2}(D)^2 + (1-Y)\frac{1}{2}\{max(0, 1-D)\} \quad (4)$$

where D is the similarity score obtained using eq. (1) and Y is the similarity label.

5 EXPERIMENTS

The aim of these experiments is to evaluate the performance and impact of different components in our pipeline. We also show how the Sentence Simplification component improves the quality and quantity of triples extracted. We perform following four experiments:

5.1 Automatic Evaluation

The aim of this experiment is to evaluate our Predicate Mapping model. Since there is no gold standard dataset for evaluation, we build a testable dataset by setting a ground truth. Consider a text triple, say R_t =<S_t; P_t; O_t>, which after performing Entity Mapping would become R_t =<S_{db}; P_t; O_{db}> where S_{db} and O_{db} are the corresponding DBpedia URIs. Now, if there is only one predicate, say P_{db}, between S_{db} and O_{db} in DBpedia then P_{db} is the ground truth for P. Such R_t constitute our test dataset.

This experiment is performed using 140,000 randomly selected Wikipedia articles. A total of 3,271,660 triples were extracted and the ground truth was established for 58,842 triples. We compare our system's results with a baseline model defined below[7].

In the baseline model, we use cosine distance metric to calculate the similarity score between a text predicate and a DBpedia predicate. Let a text predicate (P_t) be a sequence of $\{t_1, t_2, ..., t_n\}$ words. We encode P_t as the sum average of GLoVE vector embeddings of the $t_i's$. Similarly a DBpedia predicate (P_{db}) is encoded as a sum average of GLoVE embeddings of $\{d_1, d_2, ..., d_m\}$ words. The cosine similarity score between these two encodings - P_t and P_{db} is defined as:

$$CosineSimilarityScore = \frac{P_t \cdot P_{db}}{||P_t|| \cdot ||P_{db}||}$$

Table 2 shows comparative results using cosine metric baseline, the rule-based approach [10], and our model trained on SNLI dataset

[6] https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10

[7] We could not compare our results with the state-of-the-art paper ([19]) due to ambiguities in the implementation details of their paper

	Recall	Precision	F1-score
Cosine	0.2561	0.3478	0.2949
Rule-based	0.3514	0.4627	0.3994
SNLI trained - Our approach	**0.6069**	**0.7684**	**0.6781**
STS trained - Our approach	0.5382	0.6762	0.5994

Table 2: Automatic Evaluation results

[5] and SemEval STS dataset [18]. We observe that training our model on SNLI dataset gives the high *F1 score of 0.678*. It is interesting to note that the model is trained on text similarity dataset to identify the similar DBpedia triples (essentially verb phrases).

It took approximately *35 hours* to complete the steps before Predicate Mapping for 140,000 Wikipedia articles, a majority of which is network latency time due to large number of SPARQL [33] queries in defining *Candidate Set* for each triple. For the final step of Predicate Mapping, we were able to generate the mappings in approximately *six hours using the pre-trained model for 58,842 triples*.

5.2 Manual Evaluation

Since there is no standard evaluation metric to evaluate the quality of the Predicate Mapping step, it is necessary to perform a manual evaluation since in Experiment 1, we are able to consider a special case - when only one predicate exists between a particular subject and object. In this experiment, we randomly select 200 sentences and were able to extract 416 triples using our triple extraction step. We then map the predicates of these triples manually to DBpedia predicates. We feed this set of triples to our Predicate Mapping model and observe that we are able to map predicates with a *recall of 0.633* and with a *precision of 0.771*.

Based on our error analysis of the extracted triples, results show that the most errors i.e. *35.2% are caused due to triple extractor*. While *30.7% errors are caused from co-reference resolution step*, *18.3% due entities* and *15.8% due predicates*. The rule-based approach [10] indicates that the most errors (*46%* approximately) were caused due to wrong Predicate Mapping, making this step a severe drawback. *Our Predicate Mapping along with the Sentence Simplifier is able to reduce this bottleneck leading to improvement of 0.22 in F1 score.*

5.3 Sentence Simplification Evaluation

Previous studies show that significant amount of errors are due to limitations of triple extraction, especially, when dealing with complex sentences. Hence in this experiment, we aim to quantify the extent to which the Sentence Simplification model improves the triple extraction performance, when dealt with complex sentences. We analyze OLLIE information extractor performance when it is fed with i) complex sentences only ii) simplified sentences using the Sentence-Simplification idea [28]. We randomly choose 100 complex sentences from Zhu dataset [40] and obtain their corresponding simple sentences and run OLLIE triple extractor on these simple sentences. We then manually classify each triple in three of the following categories: (a) *Correct triple* is a relation triple that can be justified as true, given the sentence, (b) *Incorrect triple* is a relation triple that can not be justified as true, given the sentence, and (c) *Misleading triple* is a triple which is neither correct nor incorrect but possess incomplete information, and may lead to incorrect relation if added in the KG.

Sentence Type	Correct	Incorrect	Misleading	Total
Complex	109	40	64	213
Simple	171	33	54	258

Table 3: Sentence Simplification Evaluation results

As shown in table 3, there is not only increase in total number of relation triples extracted, but also a *36.25% increase in the number of correct triples* obtained and a *17.5% decrease in the number of incorrect triples* from OLLIE when fed with simplified sentences vs complex sentences. This entails that the simplification component helps construct larger KG with less ambiguity and less incorrect relationships.

	Accurate	Additional	Total Triples	Accuracy
History	1895	278	2173	0.872
Science	948	217	1165	0.813

Table 4: Redundancy Reduction Evaluation results

5.4 Redundancy Reduction Evaluation

The aim of this experiment is to evaluate how well our system is able to match similar relationships and entities into a homogeneous set of URIs in order to reduce redundancies in the KG. We take a Wikipedia article, say A, and paraphrase it to A'. We combine A and A' into a single document, say B. The idea is, this B will give out redundant triples since same information is paraphrased and added to it (in the form of A and A') and we want to evaluate how well these redundant triples are resolved by our system. Ideally, the triples generated from A should be identical to the triples generated from B. We perform our end-to-end KG generation system over A and B independently and study the (a) number of *accurate triples* - triples which are identical across A and B, and (b) number of *additional triples* - which are present in B but are absent in A.

We use an online paraphrasing tool called spinbot[8] to paraphrase an article. We collect 1,000 such A's and build their corresponding B's separately for two different topics of Wikipedia - Science and History. Some examples of Science topic are Chlorophyll, Photosynthesis, etc and History topic are French Revolution, World War, etc. We obtain a total of 2173 triples from History articles and 1165 triples from Science articles. We define accuracy for this experiment in the following manner:

$$Accuracy = \frac{n(a)}{n(a) + n(b)}$$

where $n(a)$ and $n(b)$ represent the number of *accurate triples* and *additional triples*, respectively. This accuracy will evaluate to 1.0 when the paraphrasing module and our system perform perfectly. The results show that, the average accuracy over Science and History articles is 84.3%. As shown in table 4, we observe that the

[8] www.spinbot.com

accuracy for History articles is 87.24%, slightly higher than *81.37% for Science articles*. We randomly sample 50 articles each of Science and History to qualitatively analyze the difference in accuracy and following are the observations:

- Science articles have a lesser share of facts but a lot of complex definitions which our Sentence Simplifier is often not able to correctly simplify. This could be improved by using a labeled training dataset consisting of complex scientific definitions and their corresponding simplified definitions for learning.
- History articles are almost entirely fact driven. This leads to high number of total triples as well as less ambiguity due to paraphrasing.

6 CONCLUSION

Most organizations have high number of textual documents that need to be processed. But domain specific document processing has not seen a lot of work and different kinds of metric are needed to evaluate such a system. In this work, we present an end-to-end system for construction of Knowledge Graph from unstructured text. Our major focus lies in:

(1) Improving the Predicate Mapping step for greater searchability in applications such as question answering and information retrieval.
(2) Developing a pruning strategy to make our KG generation system scalable.
(3) Studying the impact of the Sentence Simplification component in improving the quality of open Information Extraction techniques and evaluating the redundancy reduction in different domains of Wikipedia. Furthermore, these two experiments are novel and reproducible.
(4) Evaluating the quality of constructed KG by performing automatic and manual experiments

Even though text data from different sources have vocabulary gap, our system is able to construct a Knowledge Graph with a *F1-score of* 0.678 as shown in table 2.

As part of ongoing work, we will construct domain specific Knowledge Graphs for health and election. We will also develop a Question Answering system based on Knowledge Graphs.

REFERENCES

[1] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 344–354.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[3] Isabelle Augenstein, Sebastian Padó, and Sebastian Rudolph. 2012. Lodifier: Generating linked data from unstructured text. In *Extended Semantic Web Conference*. Springer, 210–224.

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 1247–1250.

[5] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[6] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning.. In *AAAI*, Vol. 5. Atlanta, 3.

[7] Roldano Cattoni, Francesco Corcoglioniti, Christian Girardi, Bernardo Magnini, Luciano Serafini, and Roberto Zanoli. 2012. The KnowledgeStore: an Entity-Based Storage System.. In *LREC*. Citeseer, 3639–3646.

[8] Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 355–366.

[9] Amira Abd El-atey and Sherif El-etriby. 1911. Semantic Data Extraction from Infobox Wikipedia Template. *Coordinates* 30 (1911), 261.

[10] Peter Exner and Pierre Nugues. 2012. Entity extraction: From unstructured text to dbpedia rdf triples. In *The Web of Linked Entities Workshop (WoLE 2012)*. CEUR, 58–69.

[11] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1535–1545.

[12] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*. Edinburgh, Scotland, UK.

[13] James Fan et al. 2010. Prismatic: Inducing knowledge from a large scale lexicalized relation resource. In *Proceedings of the NAACL HLT 2010*. ACL, 122–127.

[14] Jenny Rose Finkel et al. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. ACL, 363–370.

[15] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *null*. IEEE, 1735–1742.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science* 313, 5786 (2006), 504–507.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 17–26.

[19] N Kertkeidkachorn and R Ichise. 2017. T2KG: An end-to-end system for creating knowledge graph from unstructured text. In *Proceedings of AAAI Workshop on Knowledge-based Techniques for Problem Solving and Reasoning*. 743–749.

[20] Vincent Kríž, Barbora Hladká, Martin Nečaskỳ, and Tomáš Knap. 2014. Data extraction using NLP techniques and its transformation to linked data. In *Mexican International Conference on Artificial Intelligence*. Springer, 113–124.

[21] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

[22] Yeqing Li. 2017. Research and Analysis of Semantic Search Technology Based on Knowledge Graph. In *Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on*, Vol. 1. IEEE, 887–890.

[23] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. *arXiv preprint arXiv:1805.07591* (2018).

[24] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)*.

[25] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. ACM, 1–8.

[26] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

[27] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity.. In *AAAI*, Vol. 16. 2786–2792.

[28] Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 1: Long Papers)*, Vol. 1. 435–445.

[29] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. 148–157.

[30] Feng Niu, Ce Zhang, Christopher Ré, and Jude W Shavlik. 2012. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS* 12 (2012), 25–28.

[31] Rohan Padhye, Debdoot Mukherjee, and Vibha Singhal Sinha. 2014. Api as a social glue. In *Companion Proceedings of the 36th International Conference on Software Engineering*. ACM, 516–519.

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[33] Eric Prud, Andy Seaborne, et al. 2006. SPARQL query language for RDF. (2006).

[34] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[35] Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the ACL*. ACL, 73–76.

[36] Michael Schmitz et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL, 523–534.

[37] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 697–706.

[38] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep reasoning with knowledge graph for social relationship understanding. *arXiv preprint arXiv:1807.00504* (2018).

[39] Aaron Steven White and Reisinger. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1713–1723. https://aclweb.org/anthology/D16-1177

[40] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 1353–1361.

How New is the (RDF) News?

Assessing Knowledge Graph Completeness over News Feed Entities

Tomer Sagi
tsagi@is.haifa.ac.il
University of Haifa
Haifa, Israel

Yael Wolf
ywolf1@staff.haifa.ac.il
University of Haifa
Haifa, Israel

Katja Hose
khose@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

Linked Open Data and the RDF format have become the premier method of publishing structured data representing entities and facts. Specifically, media organizations, such as the New York Times and the BBC, have embraced Linked Open Data as a way of providing structured access to traditional media content, including articles, images, and video. To ground RDF entities and predicates in existing Linked Open Data sources, dataset curators provide links for some entities to existing general purpose repositories, such as YAGO and DBpedia, using entity extraction and linking tools. However, these state-of-the-art tools rely on the entities to exist in the knowledge base. How much of the information is actually new and thus unable to be grounded is unclear. In this work, we empirically investigate the prevalence of new entities in news feeds with respect to both public and commercial knowledge graphs.

KEYWORDS

Knowledge graph extension; Emerging entities; RSS; News

ACM Reference Format:
Tomer Sagi, Yael Wolf, and Katja Hose. 2019. How New is the (RDF) News?: Assessing Knowledge Graph Completeness over News Feed Entities. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3317702

1 INTRODUCTION

Humanity has been generating text since the dawn of civilization. Embedded within are stories of people, their actions, and experiences. Some are well known, their stories repeated and eventually coded into the canon of human knowledge as structured data. This type of information can be queried and disseminated in a variety of forms, ensuring its availability. However, most people, and everything about them, are lost to the anonymity of unstructured text. Similarly, modern texts, such as news articles, blogs, and web-pages, contain huge amounts of unstructured text, of which but a fraction is harvested and codified into structured forms, such as knowledge graphs.

Knowledge graphs (KG) have evolved on the World Wide Web as a standard for representing knowledge as a network of entities associated with facts and connected to other entities via properties. An increasing variety of tools to manage KG exists under the conceptual umbrella of the Semantic Web. Some KG are hand-built by domain experts, some converted from structured sources (e.g., DBpedia [2], which is created from the structured elements of Wikipedia), and some result from semi-automated extraction performed on unstructured texts, e.g., Dacura [30].

Once a Knowledge Graph is created, keeping it up to date entails monitoring its sources for new facts and emerging entities. The challenge of extracting entities and facts from unstructured text to extend existing knowledge graphs is attracting increasing attention from both industry (e.g., ambiverse.com and heuritech.com) and academic research [25, 29, 30, 34]. However, the need to ground the extracted information in entities that exist in the KG leaves many entities and predicates behind. Consider Knowledge Vault [9] (KV), widely considered state-of-the-art in this realm. KV is trained under a Local Closed World Assumption (LCWA) [9, 12], where training and testing statements are ignored if the statement does not appear in Freebase (its evolving knowledge graph). Thus, KV is trained in the absence of true new information. An empirical evaluation of the LCWA approach in [9] finds that it misses by ten percentage points when tested on data labeled without this assumption. This 10% may very well be the new information uncovered by KV! Furthermore, many of the proposed techniques [20] rely on Wikpedia and derived sources, such as Wikilinks, Wikidata, DBpedia, and YAGO, for grounding. Wikipedia growth has essentially plateaued [37]. Thus, requiring new entities and facts to be grounded in existing KG limits accepted information to high-profile people, places, events, and products on which some knowledge already exists.

Concurrently, Linked Open Data and the RDF format have become the premier method of publishing and representing structured data representing entities and facts. Specifically, media organizations such as the New York Times[1] and the BBC[2] have embraced Linked Open Data as a way of providing structured access to traditional media content, including articles, images, and video. To ground RDF entities and predicates in existing Linked Open Data sources, dataset curators provide links for some entities to existing general purpose repositories such as WordNet and DBpedia. However, much of the information is actually ungrounded, and media organizations do not regularly contribute this ungrounded information back to the general purpose repositories.

Despite efforts specifically targeting news [13], which provide a tool to extract new entities from unstructured RSS news feeds, very little research targets these missing/emerging entities. A recent survey [29] stressed the scarcity of knowledge extension approaches

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317702

[1] https://developer.nytimes.com/
[2] https://www.bbc.co.uk/ontologies

that identify and attempt to complete missing entities and not only facts about existing entities. In the latest round of the DBpedia challenge, where participants are required to locate missing information to be added to DBpedia, the winner (Matteo Cannaviccio, Roma Tre University) professed to be missing a good ground truth for measuring against and reported recall (coverage) of only 23% on a limited set of simple phrases, such as âĂIJ[Person] married [Person]âĂİ.

The limited amount of work targeting knowledge extension and the absence of datasets and evaluation methods for KG extension tools is puzzling. At face value, we cannot rule out the possibility that it indicates the absence of a problem. Perhaps the number of new entities in news streams is very low and established knowledge graphs are doing a good job of collecting them? A hint to the magnitude of the problem can be found in [13]. The authors created a manual golden standard for entity disambiguation by selecting a random 1 % of their diverse RSS feed corpus and from the resulting 70K sentences selecting only 479 entity pairs from more prevalent relation types. On this small sample, which was slightly skewed towards more popular entities by the selection method, only 456 of the 934 entities were found in DBpedia. However, it is unclear, how representative this small sample is with respect to the prevalence of new entities in the whole corpus. Furthermore, additional questions regarding the prevalence and characteristics of new entities and facts in the news remain unanswered.

To answer these questions, in this work we empirically investigate the prevalence of new facts and new entities in news articles with respect to large established public and commercial knowledge graphs. We thereby provide insight on the following questions by evaluating them over articles collected from popular RSS feeds:

(1) What percentage of the entities in news feeds is available in a public knowledge graph close to its publication?
(2) How does this figure change over time?
(3) What types of entities are better captured and which types are less represented and how does this reflect on the prevalent methods of knowledge graph extensions?

The remainder of this paper is structured as follows. In Section 2 we describe the problem of knowledge graph extension and the major techniques in use in this realm. In Section 3 we describe related domains and their relation to this work. Section 4 describes the methods we use and justify some of the methodological choices made. We then present the results of the analyses performed (Section 5), and discuss possible explanations and the implications of these results.

2 BACKGROUND

RDF Knowledge Graphs (KG) have become the de-facto standard for representing structured knowledge on the Web. An increasing variety of tools to manage and extend KG exists under the conceptual umbrella of the Semantic Web [3]. KG are based on entities, representing people, places, and abstract things, such as events. KG also contain facts describing entities through predicates. Predicates can be used to form meaningful relations between entities as well. For example, consider Figure 1, representing a small knowledge graph about two persons married to each other. The two entities *BO* and *MO* are connected via the predicates *sc:givenName* and *sc:familyName* to literals describing them and to each other via predicates using *sc:spouse*. The emergence of web ontologies has allowed further conceptualization of knowledge graphs. RDF Schema (RDFS) and the family of Web Ontology Languages (OWL) provide the ability to recognize groups of entities as a class, create *sameAs* links between entities, create *equivalentClass* links between classes, denote different predicates as *subProperties* of a general property, and additional entailment mechanisms. For example, the entities in the small example described above are instances of class *Person* maintained by schema.org[3].

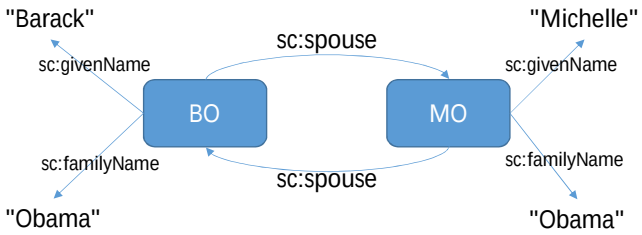


Figure 1: A Small Example Knowledge Graph

These mechanisms have allowed creating one of the most substantial achievements of modern information management, namely, the Linked Open Data (LOD) cloud, a network of KG from a diverse set of domains containing over 201 billion facts[4]. Its existence enables knowledge-based artificial intelligence services such as question answering and reasoning. Such open KG lower the entry barrier for knowledge-based application vendors by giving them access to information in context on a scale previously available only to the largest of corporations, such as Google, Microsoft, and Yahoo. Furthermore, public access to information is currently mediated by search engines that provide mostly links to websites containing the requested information rather than direct answers to questions and information in its context. Large-scale KG representations of both historic and current texts allows unmediated access to information within its factual context rather than within a web page designed by a commercial entity.

Some KG are hand-built by domain experts, some converted from structured sources (e.g., DBpedia), and some result from (semi-)automated extraction processes performed on unstructured texts (e.g., T2KG [19] and Dacura [30]). Of the three methods, the latter (extraction from text) holds the greatest potential. A report for The Economist, Cukier[5] estimated that over 95% of data in existence is unstructured.

The process of extending a knowledge graph by extracting entities and predicates from unstructured text is a complex one. As an example, consider the following article published by the Time Magazine feed[6].

Example 2.1. The Oldest U.S. Military Survivor of the Pearl Harbor Attack Has Died at the Age of 106 Ray Chavez, 106, had been battling pneumonia.

[3]https://schema.org/Person, retrieved February 1st 2019
[4]Derived from https://lod-cloud.net/, retrieved January 31st, 2018
[5]http://www.economist.com/node/15557443, retrieved July 16th, 2018
[6]http://feeds.feedburner.com/time/topstories, retrieved November 22nd, 2018

Figure 2 divides the process into three levels. On the linguistic level, information extraction techniques are used to convert text to entities and predicates (facts). Named Entity Recognition (NER) identifies entities (e.g., U.S. Military, The Pearl Harbor Attack, and Ray Chavez) followed by Coreference Resolution, which identifies other mentions of these entities (e.g., The oldest U.S. Military Survivor referring to Ray Chavez). Relation/predicate extraction attempts to attribute facts to extracted entities or create relations between them (e.g., Ray Chavez has died and is a survivor of The Pearl Harbor Attack).

The factual level has also been termed *Knowledge Graph Extension* where extracted entities are disambiguated against the existing knowledge graph. This process has also been called *Entity Linking*. If the entities are not linked to an existing KG entity with sufficient confidence, the system can attempt to predict whether or not they are new/emerging entities and thus should be inserted into the graph. The process of entity extraction and linking (EEL) is often used to annotate entities within unstructured texts with no intent of extending a knowledge graph. Similarly, extracted predicates between known entities can be predicted by comparing them to known facts involving the associated entity/entities in the KG. The problem of relation predication can be associated with the socially prominent, emerging field of fact checking. With the increasing ubiquity of self publication platforms (e.g., blogs), social networks, and new-media organizations, there is growing concern regarding the validity of facts. Identification of false information portrayed as a fact has become an active research area [7, 10, 28] but is not the focus of this work. The complete process of identifying emerging entities and new predicates was also referred to as *KG completion*.

Although in this work we focus on the factual level, for completeness we describe the conceptual level as well. On the **conceptual** level, *Ontology Learning* processes (e.g., [24]) try to generalize a collection of entities into a conceptualization of the domain. For example, after encountering multiple statements about entities declaring them to be democratic senators, we may deduce that *Democratic Senator* is a class. Similarly, the fact that Democratic senators and people share many traits but that some traits are unique to both classes can be used to derive candidate properties, adding to the evolving ontology. These candidate classes and properties can be aligned with an evolving ontology or with an external ontology using ontology alignment techniques [17].

3 RELATED WORK

Detecting a novel document in a stream of documents [18] is a task related to information retrieval (IR). However, instead of identifying novel entities or facts, existing work rather identifies documents containing novel information. Derczynski et. al. [8] evaluate NER tools on twitter tweets, which are comparable to RSS feed articles in terms of length. However, their evaluation focused on the performance of the tools on this dataset rather than the actual existence of entities in the knowledge graph and how this changes over time.

Open information systems, such as Reverb [11] and OLLIE [26], do not use a fixed ontology to classify extracted entities and predicates, but rather identify entities and facts and extract a common ontology. However, the information extracted is not linked to an existing knowledge graph. Similarly, other approaches identify information in structured sources, such as Wikipedia info boxes or other structured HTML elements (YAGO [16], DBpedia [2]) using a fixed ontology. While these tools do create knowledge graphs, they actually recreate the graph each time they are run rather than extend an existing graph. Conversely, approaches, such as NELL [6], DeepDive [32], and Google's Knowledge Vault [9], attempt to ground the information in existing KG.

All of the above mentioned systems use techniques that have been studied and evaluated separately as well. Information extraction technologies pioneered in the NLP domain, and have recently been increasingly used in conjunction with Semantic Web technologies and use cases [25, 29]. As described in the previous section, entity extraction and linking (EEL) is the process of identifying entities in unstructured text and linking them to existing knowledge bases. A wealth of work on this subject exists; recent advances include features using neural networks [20] and word embeddings [22], which can be tailored to better support events [31]. A common approach is to split this process into two distinct steps, the first being named entity recognition (NER) [1], also known as Mention Detection, and subsequently disambiguating the detected entities against a knowledge base. Approaches to perform the latter are called either entity disambiguation (ED) or entity linking (EL) [38]. However, it is those entities that are not linked/disambiguated that interest us the most in this paper, and these often do not even appear in the golden standard against which these techniques are measured.

Discovering new entities to be added to a KG is a type of knowledge graph refinement. It has also been called emerging entity (EE) discovery or out-of-knowledge-base (OOKB) entity discovery. In a recent survey, under the completeness category, Paulheim [29] has listed the limited amount of work on the subject. Notable work includes that of Hoffart et. al. [15], which attempts to identify emerging entities to be added to a knowledge graph focusing on the case where the names of new entities are ambiguous with existing KG entities. Singh et. al. [33] expand this work by providing a human-in-the loop system to identify and merge these entities to an evolving knowledge graph. Recently, Brambilla et. al. [5] have shown how using seeds of existing people in a curated set of tweets one can identify more examples of emerging entities, specifically people. However, all of these efforts have not provided an empirical analysis of the rate and composition of emerging entities in general and specifically in news streams.

4 METHODOLOGY

4.1 The News

The form in which people keep up with current events has changed over the years. Print as a primary source of news has long since been complemented, and in many ways, replaced with visual media and Internet-based sources. Moreover, even within electronic news consumption, traditional websites are being replaced with social-media based news sharing [21]. For the purposes of this work, we return to RSS feeds, used in previous work on entity extraction from news articles [13]. Although their readership is declining[7] and while most people may receive their news in alternate forms, they

[7] https://www.howtogeek.com/164375/why-google-reader-died-4-alternatives-to-rss-readers/

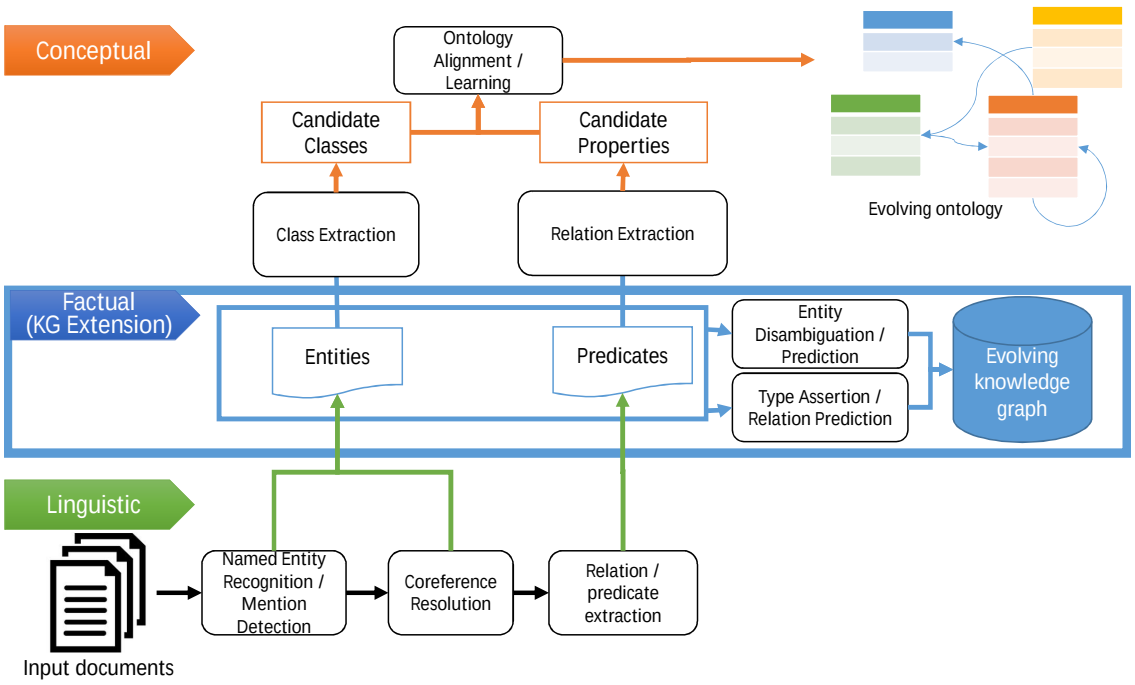


Figure 2: Knowledge extraction process

still represent a reliable method to stream news articles by topic and represent the same news being distributed by other means. Specifically, we took all functioning feeds listed by feeder[8] (one of the most popular RSS feed readers) as the most popular feeds among their users. The 91 feeds cover a wide range of interests and are categorized into the following topics: *News (18), Sports (13), Technology (20), Business (14), Politics (10), and Gaming (16).* Among these, the *News* category is the most diverse, often divided into sub-topics, some of which may overlap with the main topics listed above. We limit ourselves to news articles from English speaking countries since we are comparing against the English language versions of the knowledge graphs, which in general are more up-to-date and extensive than their counterparts in other languages.

4.2 Knowledge Graphs and Entity Linking

In this work, we examine the rate in which new entities appear in up-to-date knowledge graphs (KG) with a publicly accessible API. Table 1 was compiled from previous research [25, 29, 40] and lists known KG considered for this work. Our inclusion criteria were for a KG to be publicly accessible through an API that allows entity linking, and be updated frequently. The knowledge graphs considered are a blend of commercially driven and public. The commercial graphs considered were Google [14], Yahoo[9], and the Microsoft Concept Graph[10]. The public graphs considered were YAGO [36], DBpedia [2] – accessible through Spotlight [27], Freebase [4], Wikidata [39], and Sunflower [23].

Some notable exclusions are: FreeBase, on which a substantial number of entity linking tools (EEL) were tested over the years, was excluded due to the fact that its last public release was in 2012 and that it is now the basis for Google KG. YAGO is rebuilt from its sources using a lengthy process, which is performed infrequently. The YAGO download page[11] dated the latest version on September 20th, 2018 to June 18th, 2017. Wikidata is in many respects equivalent to DBpedia, we therefore chose DBpedia due to ease of use of the DPpedia spotlight API with respect to its Wikidata counterpart Ask Wikidata[12], which is designed to answer natural language questions but can be abused by inputting article text (see discussion on the limited amount of tools for data retrieval from Wikidata by Spitz et. al. [35]).

Our final selection consists of a commercial graph, namely Google KG (GKG) [14] and a public graph namely DBpedia [2]. For both graphs, a public API is provided and both are kept up-to-date through frequent updates from their sources. We employ the DBpedia Spotlight API[13] to parse our articles comprised of the article title concatenated with the article text. Since the Google KG API [14] receives an entity mention and returns a list of matching entities, we perform entity extraction using the Google NLP entity API[14] prior to submitting the mention to the Google KG API.

[8]https://feeder.co/knowledge-base/rss-content/rss-lists-popular-and-useful-rss-feeds/
[9]https://developer.yahoo.com/contentanalysis
[10]https://concept.research.microsoft.com/

[11]https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/
[12]https://tools.wmflabs.org/bene/ask/, retrieved February 28th, 2019
[13]https://github.com/dbpedia-spotlight/dbpedia-spotlight
[14]https://cloud.google.com/natural-language/docs/analyzing-entities

KG	Commercial?	EL API	Inclusion Decision
DBpedia	No	Via Spotlight	Include
Wikidata	No	Ask Wikidata (partial match)	Exclude
Freebase	No	Unavailable	Exclude, not maintained
Google KG	Yes	Google NLP & KG API	Include
Microsoft Concept Graph	Yes	Limited, class only	Exclude, limited API, no update policy
Sunflower	No	Unavailable	Exclude, not maintained
Wikipedia/Wikilinks/Wikidata	No	SPARQL	Exclude, same source as DBpedia
Yahoo KG	Yes	Rate limited	Exclude - rate limited
YAGO	No	Available	Included in separate evaluation.

Table 1: Knowledge Graphs

4.3 Evaluation procedure

In this section, we detail our evaluation procedure. A companion GitLab repository[15] contains the scripts and tools used to produce the presented results.

4.3.1 Preprocessing. The following preprocessing procedure is performed on all articles. HTML tags and links are removed if present, HTML ASCII codes are converted back to their corresponding utf-8 symbols (e.g. … is converted to …). Article text is truncated at 3200 characters which covers the length of 96.45% of the articles to avoid overloading our human annotators.

4.3.2 Human Annotation. After extracting the entities and looking them up in the KGs we use the *figure eight*[16] crowd sourcing platform for human tagging using the following procedure. A task is generated for each article where the text is annotated using square brackets for the discovered entities. The annotator is tasked with validating that all entities were identified. For example, consider the task shown in Figure 3. An article describing a couple of known local celebrities renovating their home. The entity extraction and linking process against DBpedia correctly identified the Sydney suburb of *Double Bay*, model Jesinta Franklin and the AFL (Australian Football League). However, Buddy Franklin, the AFL player, was misidentified as Benjamin Franklin, the American polymath. After marking "Missing/wrong entities...", a more detailed list of options appears allowing the annotator to list the missing entities and the wrongly identified entities. Tasks flagged for missing entities were manually verified by using the GKG and DBpedia public search interfaces. If they were indeed missing, they were added to a list of missing entities. In the example presented, Buddy Franklin actually exists in DBpedia (and GKG) and was just improperly disambiguated by DBpedia spotlight.

The lookup process and tagging procedure are repeated after one month, and two months to examine changes over time, specifically, which novel entities were added to the knowledge graphs and which remain excluded.

5 RESULTS

A total of 866 cases of missing entities were found out of 13,456 named entities detected by the named entity recognizers. Of these, 378 were missing from DBpedia, 488 were missing from GKG. Figure 4 details the overlap between these groups. For example, of the 378 missing from DBpedia, 81 were missing only from DBpedia and existed in GKG. Similarly, Of the 488 reported above, 191 were missing only from GKG. Note that 297 entities were missing from both DBpedia and GKG, and therefore counted twice in the total number reported above.

After removing duplicates, 408 unique missing entities were left. We classified the entities into six classes according to *schema.org*[17] top level classes, namely People, Products, Creative Works, Organizations, Events and Health & Medical. Table 2 presents the number of entities missing by class with a few demonstrative examples of each class.

To better understand the reasons of omission, we further refined two of the classes described above, namely People and Creative Works. In the People category, missing entities were distributed as follows: 43 journalists, 21 criminals, 17 victims, 20 artists, 31 functionaries in companies, eight sports people, five relatives of famous people, four public functionaries, four researchers, three suspects, three fictional characters, and 37 "others". We speculate that the reason most people are missing from a graph can be attributed to the methods employed by EEL in KG extension, i.e linking new elements to existing ones. Thus, random people (such as the subcategory "others") who appear in a news article will probably not be added to a KG. For example, victims (and criminals) of isolated attacks are less likely to be added than a mass shooter. Having said that, we observed that GKG did contain some entities of victims and criminals of such attacks. The second category, Creative Works was distributed as follows: 33 games, 16 art pieces, three web content, one website, and one "other". Among the missing games, there were old ones, new games released recently, new releases in a series, and games with an upcoming release date which may be the reason why they were not in a KG yet. The art pieces contained mostly characters of animated films that may have a limited number of followers in monitored sources such as Wikipedia, preferring their own wiki-type constructs such as Fandom/Wikia[18] .

In addition, we noticed an interesting trend regarding products. Some companies' products did not appear in one of the KG, mainly Apple and Google products, which were only in DBpedia. This

[15] https://gitlab.com/ts.tomersagi/newnews
[16] https://figure-eight.com
[17] https://schema.org/docs/schemas.html, retrieved Jan. 15th 2019
[18] https://www.wikia.com/explore

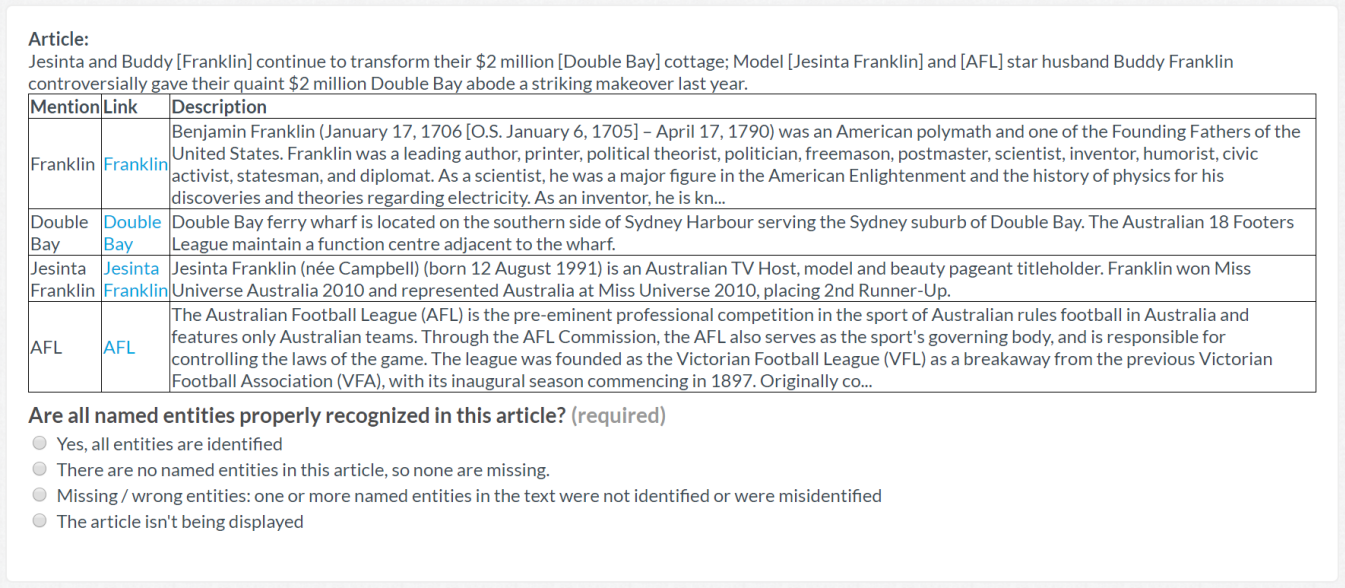


Figure 3: Example of Figure Eight Task

Class	Number	Examples
People	196	Josh Edelson, a photographer and Yasha Haddaji, Nintendo RussiaâĂŹs General Manager, who are both missing from GKG and DBpedia
Products	109	Recon 200, a new headset by Turtle Beach, missing from both GKG and DBpedia and PlayStation 4, video game console, missing from GKG
Creative Works	54	"A Head Of A Young Man ", a painting by Peter Paul Rubens, which is missing from DBpedia, and AssassinâĂŹs Creed Rebellion, a game, which was missing from both GKG and DBpedia
Organizations	41	Chime Banking and Empower, both banks missing from both GKG and DBpedia
Events	7	Assembly Elections in India which was missing from GKG and FortniteâĂŹs Winter Royale Online Tournament, which was missing from both GKG and DBpedia
Health & Medical	1	Acute Flaccid Myelitis, a rare disease, missing from GKG.

Table 2: Classes and Examples

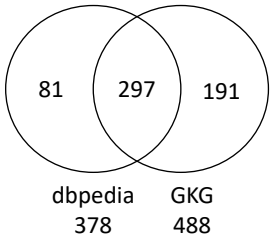


Figure 4: Venn Diagram Depicting Missing Entities

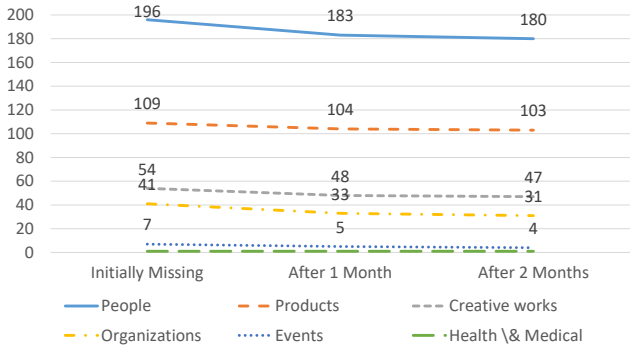


Figure 5: Missing Entities over Time (by class)

could result from company policy but the real reason is unknown to us.

After one month, we reexamined the above unique entities to see which were added to the knowledge graphs. A total of 34 additions were found: eight to DBpedia and 26 to GKG. Six creative works were added: four of them were added to GKG – all video games, three are new, and one has a future release date (February

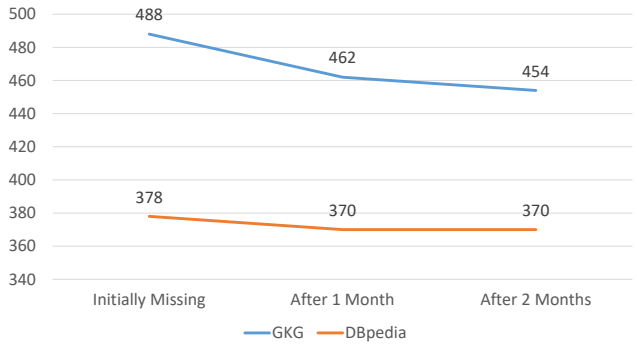


Figure 6: Missing Entities over Time (by graph)

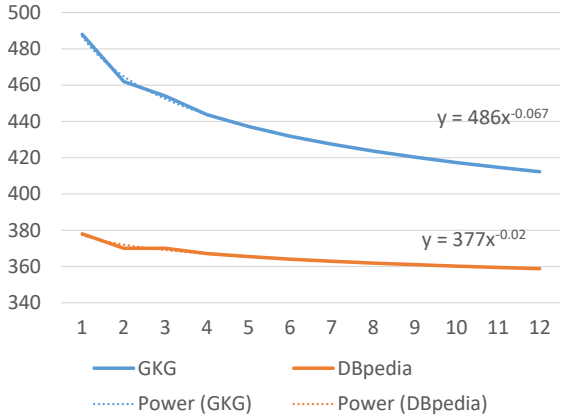


Figure 7: Extrapolated Reclamation Rate

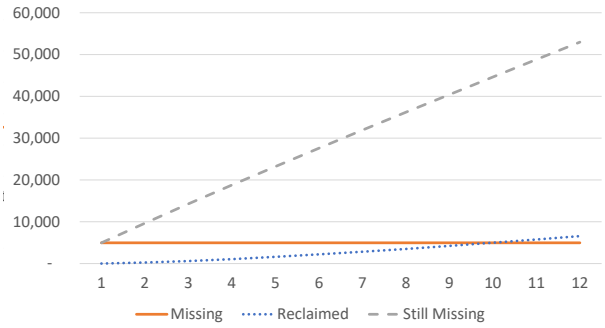


Figure 8: Cumulative Effects of Reclamation Rate

2019) and is part of a series; two were added to DBpedia, the latter game and a fictional character in an animated film. 13 people were added: two to both knowledge graphs, one of which is a young basketball player; seven to GKG including three victims; two to DBpedia, a young soccer player and an up-and-coming actress. Eight organizations were added: one to both knowledge graphs, a financial services company; six to GKG, including a primary school in London. Five products were added: one product was added to both knowledge graphs, and the rest to GKG. Two events were added to GKG, both wildfire events in the United States. After an additional month, we repeated the procedure. Only eight entities were added. All were added to GKG: Three people – young soccer players, two organizations, one event, one creative work and one product. Figures 5 and 6 show the number of missing entities over time by class and graph respectively.

To better understand which types of news streams the missing entities come from, we return to the original feed classification. As mentioned above, feeds were classified at the source into one of five categories, namely Gaming, News, Politics, Sport, and Technology. We manually validated this voluntary classification and found a few articles classified as News but that better fit one of the other categories. Table 3 presents an analysis of the base rate of missing entities in different categories. As is evident, the domains of Gaming and Technology feature a higher base rate than News and Sports. While it is tempting to conclude that there are no new entities in Politics, the small number of articles and the fact that our examination period was between election cycles preclude such a conclusion.

6 CONCLUSIONS AND FUTURE WORK

So... how new is the news? Our empirical analysis found that on average, about six percent (6%) of entities mentioned in a variety of news feeds were missing from the knowledge graph. While the rate may sound small, one must consider that only a small percentage of these may eventually be recovered. To assess the extent of eventual recovery we extrapolated (Figure 7) the number of recovered entities to 12 months by fitting inverse power functions to the original recovery rates. The fitted functions are displayed next to the extrapolated figures. Both functions represent a decaying rate of recovery which best fits the observed data ($R^2 > 0.86$ for DBpedia and $R^2 > 0.98$ for GKG). The overall number of reclaimed entities is projected to be 76 for GKG and 19 for DBpedia, representing 15% and 5% of the missing entities respectively. Figure 8 presents the result of applying this rate and extrapolating the number of newly missing entities to a 12 month period. Under a conservative assumption of a stable rate of new missing entities, it is obvious that the rate of reclamation is vastly out-paced by the rate of new entities. Thus, knowledge about an increasing number of people, products, and creative works is lost over time.

As mentioned above, there is a scarcity of datasets and evaluation methods for identifying and linking novel entities. The work presented in this paper highlights the need for both a dynamic and a static evaluation method for knowledge graph extension. A dynamic method should challenge submitted tools with new information that cannot be found in public knowledge graphs and the static dataset and evaluation protocol should use news articles collected recently with a strictly enforced temporal separation between this evaluation dataset and the data sources on which evaluated systems rely on.

In future work, we intend to explore the underlying assumptions of existing emerging entity identification systems to uncover the methodological reasons for this growing problem. Furthermore, we wish to validate our extrapolation assumptions by scaling this work over time and into additional types of news sources.

Feed Category	Cases of Missing Entities	Total Articles	Missing entity/article	Extracted entities	Missing entities/extracted
Gaming	224	214	1.0	2,786	8%
News	251	557	0.5	4,566	5%
Politics	0	14	0.0	113	0%
Sport	50	252	0.2	2,759	2%
Technology	341	309	1.1	3,232	11%

Table 3: Analysis of missing entity base rate

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *27th Int. Conf. on Comp. Linguistics, COLING '18*. ACL, 1638–1649.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *ISWC/ASWC*. Springer, Busan, Korea, 722–735.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284, 5 (2001), 34–43.

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase. In *SIGMOD '08*. ACM Press, New York, New York, USA, 1247.

[5] Marco Brambilla, Stefano Ceri, Emanuele Della Valle, Riccardo Volonterio, and Felix Xavier Acero Salazar. 2017. Extracting Emerging Knowledge from Social Media. In *WWW '17*. ACM Press, New York, New York, USA, 795–804.

[6] Andrew Carlson, Justin Betteridge, and Bryan Kisiel. 2010. Toward an Architecture for Never-Ending Language Learning. In *AAAI '10*. AAAI Press, 1306–1313.

[7] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A Content Management Perspective on Fact-Checking. In *WWW '18*. ACM, 565–574.

[8] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Inf. Processing and Mgt.* 51, 2 (2015), 32–49.

[9] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *KDD '14*. ACM, New York, NY, USA, 601–610.

[10] Gonenc Ercan, Shady Elbassuoni, and Katja Hose. 2019. Retrieving Textual Evidence for Knowledge Graph Facts. In *ESWC*. Springer.

[11] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP '11*. ACL, 1535–1545.

[12] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW '13*. ACM Press, New York, New York, USA, 413–422.

[13] Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2013. Real-Time RDF Extraction from Unstructured Data Streams. In *ISWC '13*, Vol. 8218. Springer, 135–150.

[14] Google. 2001. Google Knowledge Graph Search API. Retrieved Feb. 1st, 2019 from https://developers.google.com/knowledge-graph/

[15] Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *WWW '14*. ACM Press, New York, New York, USA, 385–396.

[16] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In *IJCAI '13*. IJCAI/AAAI, 3161–3165.

[17] Prateek Jain, Pascal Hitzler, Amit P Sheth, Kunal Verma, and Peter Z Yeh. 2010. Ontology alignment for linked open data. In *ISWC '10*. Springer, Springer, 402–417.

[18] Margarita Karkali, François Rousseau, Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient online novelty detection in news streams. In *Web Inf. Sys. Eng. âĂŞ WISE '13. LNCS.*, Lin X., Manolopoulos Y., Srivastava D., and Huang G. (Eds.), Vol. 8180. Springer, Berlin, Heidelberg, 57–71.

[19] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2KG : An End-to-End System for Creating Knowledge Graph from Unstructured Text. In *Proceedings of AAAI Workshop on Knowledge-based Techniques for Problem Solving and Reasoning*. AAAI Press, 743–749.

[20] Nikolaos Kolitsas, Eth Zürich, Octavian-Eugen Ganea, and Thomas Hofmann. October 31 - November 1, 2018. End-to-End Neural Entity Linking. In *CoNLL '18*. Association for Computational Linguistics, Brussels, Belgium, 519–529.

[21] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. 2015. News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks. *Social Media + Society* 1, 2 (sep 2015), 205630511561014.

[22] Phong Le and Ivan Titov. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In *the 56th Annual Meeting of ACL '18*. ACL, 1595–1604.

[23] Marek Lipczak, Arash Koushkestani, and Evangelos Milios. 2014. Tulip: Lightweight Entity Recognition and Disambiguation Using Wikipedia-Based Topic Centroids. In *1st Int. Work. on Entity Recognition & Disambiguation - ERD '14*. ACM New York, NY, USA, Gold Coast, Queensland, Australia, 31–36.

[24] Alexander Maedche and Steffen Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent systems* 16, 2 (2001), 72–79.

[25] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2018. Information Extraction meets the Semantic Web: A Survey. *Semantic Web* Pre-press, Pre-press (2018), 1—-81.

[26] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *EMNLP and CoNLL 12', Jeju Island, Korea*. ACL, 523–534.

[27] Pablo N Mendes, Max Jakob, Andrés García-silva, and Christian Bizer. 2011. DBpedia Spotlight : Shedding Light on the Web of Documents. In *7th Int. Conf. on Semantic Sys. (I-Semantics).*, Vol. 95. ACM, 1–8.

[28] Steffen Metzger, Shady Elbassuoni, Katja Hose, and Ralf Schenkel. 2011. S3K: seeking statement-supporting top-K witnesses. In *CIKM*. ACM, 37–46.

[29] Heiko Paulheim. 2016. Knowledge graph refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web* 0 (2016), 1–23.

[30] Peter N Peregrine, Rob Brennan, Thomas Currie, Kevin Feeney, Pieter François, Peter Turchin, and Harvey Whitehouse. 2018. Dacura: A new solution to data harvesting and knowledge extraction for the historical sciences. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51, 3 (2018), 165–174.

[31] Vinay Setty and Katja Hose. 2018. Event2Vec: Neural Embeddings for News Events. In *SIGIR*. ACM, 1013–1016.

[32] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. 2015. Incremental knowledge base construction using DeepDive. *pVLDB* 8, 11 (jul 2015), 1310–1321.

[33] Jaspreet Singh, Johannes Hoffart, and Avishek Anand. 2016. Discovering Entities with Just a Little Help from You. In *CIKM '16*. ACM Press, New York, New York, USA, 1331–1340.

[34] Anton Södergren and Pierre Nugues. 2017. A Multilingual Entity Linker Using PageRank and Semantic Graphs. In *the 21st Nordic Conf. on Comp. Linguistics, NODALIDA '17, Gothenburg, Sweden*. ACL, 87–95. http://www.ep.liu.se/ecp/131/011/ecp17131011.pdf

[35] Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz, and Johanna Geiß. 2016. State of the union: A data consumer's perspective on wikidata and its properties for the classification and resolution of entities. In *10th AAAI Conf. on Web and Social Media*. 88–95.

[36] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *WWW '17*. ACM New York, NY, USA, Banff, AB, Canada, 697–706.

[37] Bongwon Suh, Gregorio Convertino, Ed Chi, and Peter Pirolli. 2009. The Singularity is Not Near: Slowing Growth of Wikipedia. In *In the 5th Int. Symp. on Wikis and Open Collab. SE - WikiSym '09*. ACM, 1–10.

[38] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation.. In *IJCAI '15*. AAAI Press, 1333–1339.

[39] Denny VrandeÄŒiÄĆ and Markus KrÃűtzsch. 2014. Wikidata: A Free Collaborative Knowledge Base. *Commun. ACM* 57 (2014), 78–85.

[40] Gerhard Weikum, Johannes Hoffart, and Fabian Suchanek. 2016. Ten Years of Knowledge Harvesting : Lessons and Challenges. *IEEE Data Eng. Bull.* 39, 3 (2016), 41—-50.

Trust and Privacy in Knowledge Graphs

Daniel Schwabe
Department of Informatics
Pontifical Catholic University
Rio de Janeiro, Brazil,
dschwabe@inf.puc-rio.br

Carlos Laufer
Department of Informatics
Pontifical Catholic University
Rio de Janeiro, Brazil,
laufer@globo.com

ABSTRACT

This paper presents the KG Usage framework, which allows the introduction of KG features to support Trust, Privacy and Transparency concerns regarding the use of its contents by applications. A real-world example is presented and used to illustrate how the framework can be used.

CCS CONCEPTS

• Information systems → Social networks • **Information systems → Semantic web description languages** • Security and privacy → Social aspects of security and privacy

KEYWORDS

Knowledge Graph, Privacy, Transparency, Trust, Semantic Web.

ACM Reference format:

Daniel Schwabe and Carlos Laufer. 2019. Trust and Privacy in Knowledge Graphs. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317705

1 Introduction

The term "Knowledge Graph" (KG) was introduced by Google in 2012[1], although to this day there isn't a precise definition of the term [11]. Nevertheless, graph-based databases were available before this (e.g., Wordnet [23], DBPedia [19], Yago [34], CYC [20], NELL [7], and additional ones (e.g. ConceptNet [32]) continue to be created on a regular basis.

While the graph model or some variant has been used in several KGs, it has already been observed that using only nodes as the "granule" of information is too low to express complex types of information, such as events, or time-varying data. For example, Wikidata [38] is organized around Items described by a collection of Statements [12]. Another reason for having more complex "granules" is recording provenance (meta) data, which is a fundamental part of data in some domains such as life-sciences [17].

KGs differ also on the way they are built (populated). A few are curated (e.g., CYC), others rely on crowdsourced information (e.g. Wikidata); most extract information from structured, semi-structured or textual information harvested from the Web.

The multiplicity of sources and various extraction approaches naturally raises the issue of data quality and confronts the user of the data in the KG with the issue of trusting, or not, the information obtained from the KG. For some types of information, for example in case of online reviews online and social media, this trust can have a direct effect on commercial success (e.g. [2]). This highlights the fact that data, ultimately, expresses a belief, opinion or point of view of some agent.

From a broader perspective, information (and knowledge) has become the prime resource in the Third Industrial Revolution, also called the Digital Age – when digital technologies enabled new ways of generating, processing and sharing information [26] [9], and is becoming more so as we move into the Fourth Industrial Revolution (4IR) [28]. The 4IR is characterized by a fusion of technologies, which is blurring the lines between the physical, digital, and biological spheres.

Increasingly, systems and applications operate in a context which the flow of information has direct bearing on daily lives of billions of people, where two fundamental characteristics of the use of such information emerge – Transparency and Privacy.

Transparency can be seen as the quality that allows participants of a community to know what the particular processes and agents are that are being used in its functioning. It is generally regarded as a means to enable checks and balances within this community, ultimately providing a basis for trust among participants. Considering the community as the whole society, these checks and balances are reflected in its political system, to prevent misuse by any of the parties involved.

One of the mechanisms created to increase transparency in political systems is the enactment of legislation ensuring the right of its members to access to information in a variety of contexts, ranging from government-produced information and data to consumer-related information regarding goods and products, as well as the right of individuals to freely create, publish and access information.

The free flow of information, on the other hand, may conflict with another basic human right, that of Privacy [37]. There are many definitions for Privacy [24] , but in essence they all refer to the right of an individual to control how information about her/him is used by others.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317705
[1] https://googleblog.blogspot.com/2012/05/introducing-knowledge graph-things-not.html

In order to deal with the myriad of often conflicting cross-cutting concerns, Internet applications and systems must incorporate adequate mechanisms to ensure compliance of both ethical and legal principles.

In order to be effective, we claim that the *use* of Knowledge Graphs must support these concerns – trust, privacy and transparency. In this paper we propose a framework that enables this support.

2 Background Concepts

Before detailing our framework, we briefly present our definition for each of the basic concerns. We have detailed each of them in other publications, as referenced in each sub-section.

2.1 Trust

The issue of trust has been prevalent in the Internet since its popularization in the early 90s (see [14] for a survey), with a focus on the lower layers of the Internet Architecture, emphasizing authentication. More recently, with the advent of the Web and social networks, the cybersphere, and society as a whole, has become heavily influenced by information (and misinformation) that flows in news sites and social networks in the Internet. There are many studies carried out in several disciplines attempting to characterize and understand the spread of information in the cybersphere, as well as how this affects society (see [21] for an overview).

The original vision for the Semantic Web included a "Trust" layer, although its emphasis was more on authentication and validation with static trust measures for data. There have been many efforts in representing trust, including computational models - a general survey can be found in [25]; [3] presents an excellent earlier survey for the Semantic Web; and [30] surveys trust in social networks. In the Linked Data world, it is clear that facts in Semantic Web should be regarded as claims rather than hard facts (e.g., [6]), which naturally raises the issue of trust on those claims.

As proposed initially in [1] and later in [18] and [29] the approach used is based on the work of Gerck [13] and Castelfranchi et al. [8], taking the view that trust is "knowledge-based reliance on received information", that is, an agent decides to trust (or not) based solely on her/his knowledge, and the decision to trust implies the decision to rely on the truth of received or on already known information to *perform some action.*

In terms of a Knowledge Graph, an agent wishing to perform an action must first filter those information items it deems "trusted", i.e., it will use them to perform the intended action. Since it is not possible to "half-act", in this sense trust is binary – either the agents trusts the information, or it does not. A more extensive discussion can be found in [29].

Trusted information is the basis for supporting privacy and transparency, as discussed next.

2.2 Privacy

There are many definitions for privacy - see for example [31]; [24] [33] present surveys on multi-party privacy. Tavani [35] classified privacy theories into four broad types: the nonintrusion, seclusion, limitation, and control theories. For our research, we adopt the Restricted Access/Limited Control (RALC) Theory proposed by Moor and Tavani [36]. RALC presupposes that an adequate theory of privacy needs to differentiate the concept of privacy from both the justification and the management of privacy. In this sense, the RALC framework has three components: an account of the concept of privacy, an account of the justification of privacy, and an account of the management of privacy. "*RALC requires that one must have protection from intrusion and interference and information access, it addresses concerns not only about protecting informational privacy (as described in the control and the limitation theories) but also about protection against the kinds of threats described in the nonintrusion and the seclusion theories as well.*" [35].

Based on this theory, we define Privacy as *"controlled access to information related to an agent"*. In order to ensure privacy, it is necessary to answer three questions:

- Q1: What types of Actions are allowed (and controlled) over Knowledge Items (KIs)?
- Q2: What are the relation types, between some Agent and an KI, which entitle this Agent to establish a Privacy Rule governing Actions over that KI?
- Q3: How to resolve conflicts between applicable rules?

2.3 Transparency

Generally speaking, according to Meijer [22] and others [27], transparency can be defined as "the availability of information about an actor that allows other actors to monitor the workings or performance of the first actor." It contemplates the "capacity of outsiders to obtain valid and timely information about the activities of government or private organizations"[2].

Transparency presupposes the involvement of an observed and an observer [5]. In contrast to privacy, that is concerned with information about individuals, transparency concerns any type of information, although it may make a difference if the producer (author) is an organization or an individual [16].

Transparency is also related to privacy. In the so-called "Attention Economy" [10], for example, the information about users and consumers are a primary source of value, and companies actively seek to obtain as much information about users. This can be in direct conflict with privacy rights of users, who have the right to control how information about him/her is used by others. In this state of affairs, transparency can support the resolution of potential conflict with privacy policies, since the disclosure of information about the company's processes and procedures associated to this individual contributes to trusting

[2] https://www.britannica.com/topic/transparency-government#accordion-article-history

that it is indeed compliant with these regulations and thus stimulating authorization on the part of the users.

Seen in a more abstract manner, both Privacy and Transparency relate to controlling actions over information, and who can define such controls. As such, they can be regarded as two points in the same control dimension. Privacy tends to limit or restrict actions over information items, whereas Transparency tends to allow (in some cases, mandate) actions over them, which explains the natural tension the exists between the two.

3 A Motivational Example - Disaster Relief Donation

The goal of this example is to illustrate the complex and interdependent nature trust, privacy and transparency; it will later be used to show the expressive power of the framework and how it incorporates the various concerns at play.

Consider a scenario in which a disaster has occurred, and Ed wants to donate some money to help with the disaster relief actions. Ed has received several donation requests from different organizations, therefore he must choose one to make the donation. However, given his past experience, Ed wants to make sure that the donation money will actually be used for the relief actions, rather than being misused, e.g., funding the organization's basic infrastructure, or employed in another action, or even pocketed by unscrupulous officers of the receiving organization.

Ed formulates a rule that says that he only trusts organizations that openly publish who are their financial officers, and their financial records. Financial records must be validated by accredited audit organizations.

Furthermore, because of personal reasons, Ed does not want to contribute to an organization in which George is an officer. This rule can be regarded as an application-related rule, akin to what is referred to as a "business rule" in traditional software development.

Let us assume that there is a law that stipulates that not-for-profit organizations must publicly identify their officers and that George is an officer of ReliefOrg, an NGO dedicated to raising funds for and helping disaster relief efforts. George, being a very reserved person, has a privacy rule that stipulates that his association with any organization, including ReliefOrg, should not be made public.

Ed received a request for donation from ReliefOrg and needs to decide whether he should donate or not.

The first step Ed follows is to verify is the officers of ReliefOrg are published in the KG[3]. Here we can see a potential conflict between transparency and privacy rules. George's relation with ReliefOrg should be accessible in the KG, according to not-for-profit legislation. On the other hand, George's privacy rules would prevent this access. Since transparency legislation in this case has higher precedence than personal rules, this association can be accessed by Ed's rules, and therefore ReliefOrg would not be accepted as a recipient of donations by Ed.

Consider now a slightly different scenario. Ed has no personal objections to George, but wants to check the financial integrity of ReliefOrg. He retrieves the financial report for ReliefOrg from the KG, but wants to make sure it has been audited, so he looks for a certification of the financial report. He finds out it has been audited by AuditInc, who he has not heard of, leading him to verify that it has an accreditation certificate from a public authority. If such a certificate is available, Ed analyses the financial report finding nothing wrong in principle.

However, tipped by a friend, Ed learns that George may in fact be one of the owners of AuditInc. He then checks AuditInc to see if its owners are listed, and whether George is one of them. Given George's privacy rule plus the fact that AuditInc is not a not-for-profit organization, his privacy rule would prevent access to the owner relationship, so Ed would not see that George is one of the owners, thus deciding to contribute. Note that in this scenario, George's privacy rule would apply not only to AuditInc's information, but also to information furnished by others, for example, a photo on a social network where George appears in the annual Christmas party of AuditInc with a caption mentioning his role as one of the partners, or perhaps a badge on his neck identifying him as such.

We next describe our proposed KG Usage framework and subsequently analyze this example showing how it can be represented by it.

4 A Summary of the KG Usage Framework.

Figure 2 shows a diagram of the use of information within a KG. Using a KG is represented by a Request made by some Agent for an Action over a Knowledge Item (KI).

We assume the existence of an underlying RDF graph, which would be equivalent to the "traditional" definition of KGs. The actual KG is formed by defining several named graphs over this underlying RDF graphs as a way to structure the RDF triples into Knowledge Items, similarly to Items and Qualifiers in Wikidata [12].

Thus, a Knowledge Graph represents a collection of interlinked descriptions of KIs – real-world objects, events, situations or abstract concepts – where:

- Descriptions have a formal structure that allows both people and computers to process them in an efficient and unambiguous manner;
- Entity descriptions contribute to one another, forming a network, where each entity represents part of the description of the entities, related to it.[4]

We propose to represent the KG as a collection of Knowledge Items (KIs), each of which as a nanopublication[5] [15].

[3] See for instance https://permid.org or https://opencorporates.com as examples of KGs with this type of information.

[4] https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

[5] http://nanopub.org

4.1 KI Representation

A KI, as any nanopublication, comprises an *assertion graph*, a *provenance graph* and a *publication info* graph.

The *assertion* graph of a KI contains a set of assertions about its content. The assertions in this graph are a subset of the assertions in the underlying RDF graph. As an example, if a KI refers to an event, it would have statements about the participants and their roles, location information, date information, depictions (photos, videos, ...) and so on.

The *provenance* graph of the nanopublication will contain provenance information about the assertions in the assertion graph (e.g.; what image or natural language processing software was used, recorded location info, whether the assertions were inferred using some inference engine, etc.). The provenance graph can be used to represent, to the desired level of detail, the supporting information for the assertions. For example, if an automated face recognition algorithm was used, the provenance information represented in the *provenance* graph of the nanopublication. may inform which algorithm, which parameters were used in this particular case, and a confidence factor. Another use of provenance can be seen in the case of a statement stating that, for example, <Barack Obama> placeOfBirth <Hawaii>. The provenance information can include documentation to support its truthfulness, such as a reference to a birth certificate that states that indeed the place of birth of Barack Obama is Hawaii.

The *publication info* graph will contain metadata about the creation of the KI itself (as opposed to the information contained in its *assertions* sub-graph).

4.2 Controlling usage

Since both Privacy and Transparency refer to actions over some information (in a KI), an Authorization must be granted for this Request, according to the Rules set forth by stakeholders. Stakeholders include persons "related" to the KI, as well as institutional agents such as "the State" (whose rules are stated as laws). Rules may be based (drawn) on any information available in the KG. We refer to both Privacy and Transparency rules generically as Usage Rules.

Figure 1 present the basic algorithm to evaluate a Request. The first step in granting an Authorization for the action is to determine which KIs in the KG are trusted by, the Agent requesting the Action. This is realized by a Trust Engine that collects applicable rules and includes them in a Trusted sub-graph of the KG. We refer to the set of trusted KIs by Agent *a* as the Trusted KG(a).

Once the Trusted KG has been determined, a set of usage rules are evaluated, which entails determining applicable rules (UsageRuleSet), evaluating them (EvalUsageRuleSet) and resolving conflicts if they arise. In order to evaluate an usage rule, it is also necessary to evaluate the trust rules of the author of that usage rule, which needs to be based on trusted KIs. Thus, a set of Trusted KG(a_i) is defined, one for each Author a_i of an applicable usage rule. The result of this process defines the final Authorization for the Action in the Request that was made. If the Authorization is "Allowed", the Action is carried out using the TrustedKG of the Agent that requested the Action.

```
Define EvalRequest(Agent, Action, KI),
1.      Let TS <- TrustRuleSet(Agent), RS <- UsageRuleSet(KI).
2.      Let TGA <- EvalTrustRuleSet (TS, Agent, Action).
3.      Let Aut <- EvalUsageRuleSet(RS, Agent, Action, KI)
If Aut = "allowed", Execute (Action, KI, TGA).
Define EvalUsageRuleSet (RS, Agent, Action, KI}
    1. Let RS <- Sort-by-Precedence(RS, decreasing).
    2. Let A <- DefaultAuthorization.
    3. For each R in RS,
        a) Let AR <- EvalRule(R, Agent, Action, KI);
        b) If AR = "Allowed" or AR = "Denied", return AR.
    4. Return A.
```

Figure 1: Request Evaluation Algorithm

To determine the final Authorization value, a conflict resolution strategy must be employed, which is in turn subject to Governance Rules.

This algorithm abstracts the essential decisions that must be made, to wit:

1. Who can formulate a rule for a given KI? – in line 1;
2. How are conflicts between rules resolved? – in line 5-8
3. What are the allowed actions over KIs? – in line 4.

We detail possible answers to these questions in the following sub-sections.

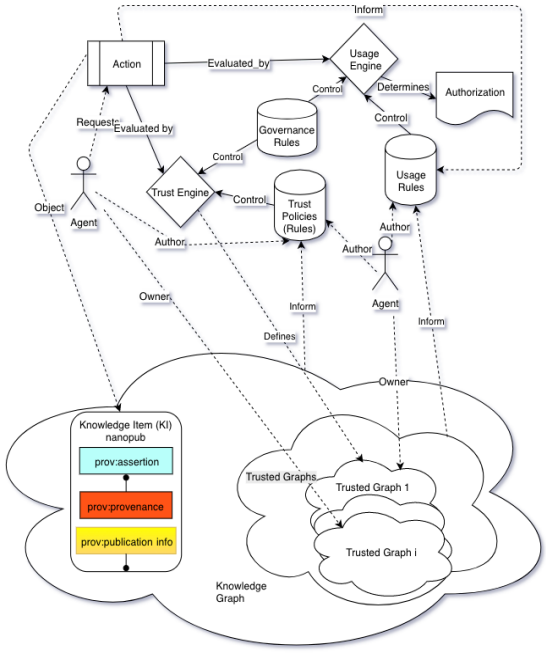


Figure 2 Usage process for information in a KG

4.3 Rules

The UsageRuleSet(KI) function determines what are the applicable rules given a KI. For trust rules, it is the author of the

request. For privacy rules this corresponds to anwering the question "who has the right to define a Privacy Rule that controls actions over this KI?" The definition of Privacy itself indicates that it must be any agent that is somehow *related* to the information contained in the KI. Any useful instantiation of the framework must spell out what are the accepted relation types, which can include:

- An identification property for any agent that is included in the KI – for example, some person appearing in a posted photo or video;
- Any relation denoting referral to a person included in the KI – for example, some person cited in a post;
- Any creation or authorship relation;
- Any agent related to the creation of the KI – for example the author of a video posted by someone else;
- Any agent who has legal jurisdiction over an agent identified or mentioned in the KI;

The Agent representing the legal system(s) that has(have) jurisdiction over the KG, over the Agent or over the Action request.

The presence of such relations can directly occur in the KG (i.e., as a typed edge), or as a composition of valid relations. Furthermore, the rules may (or may not) allow the use of inferred relations in the KG.

Rules are of the form antecedent => consequent, both of which are sets of statements [1]. The antecedent of privacy rules may refer to any statements in the KG, including

- Any statements in the graphs in the KI's nanopublication (assertions, provenance and publication info);
- The identity of the agent requesting permission;
- The type of Action;
- Information in the KG serving as contextual information), such as
- Date/time of the request;
- Current location of agents involved;

Seen as nanopublications, the actual specification of the rule is given in its assertion graph, using a notation such as N3Logic [4] or SWRL[6]. Governance Rules are Rules that include other Rules (in either of antecedent and consequent), so in this sense they are meta-rules.

4.4 Conflict resolution

The Sort-by-Precedence(RS, decreasing) function call sorts the enabled rules are sorted in descending precedence order, typically combining several types of information to establish order relations between rules.

Some possible complementary order relations that can be employed are:

- Hierarchical relations between users – rules defined by a higher-ranked user take precedence over rules defined by lower-ranked ones. For example, rules established by laws take precedence over rules stated by individuals;
- Hierarchical relations between relation types. Rules defined by users related to the KI through a higher-ranked relation type take precedence over rules defined by users related via a lower-ranked relation type. For example, one may state that "identifies" takes precedence over "mentions". Thus, in a video where a person A is identified, and person B is mentioned in a conversation, person A's rules would take precedence over person B's rules.

Since hierarchies are partial orders, they may not completely define precedence, so further conflict resolution strategies are still needed. Such and Criado [8] identified six categories of strategies that can be employed. Most require user involvement at runtime, but the aggregation-based class can be easily incorporated into an algorithm. Strategies in this class define an aggregation function such as consensus, majority, minimum fixed number of votes, permit-overrides, deny-overrides, etc. – see also [24] – and replace the set of conflicting rule results by a single aggregated result.

An alternative to the aggregation approach is to decompose the KI into finer-grained elements so that each is subject to only one rule. This makes sense when the Action to be performed can be stated as the composition of the same operation performed on each element independently. For the elements where the Action is denied, the resulting composite object is altered, such as blurring the face of a given person in a group photo.

5 Example Scenario revisited

In this section we re-examine the motivational scenario light of the KG usage framework, showing how the main important aspects are represented.

The first rule captures Ed's requirement about Non-Profits: he only trusts organizations that openly publish who are their financial officers, and their financial records. Financial records must be validated by accredited audit organizations.

The trust rule below captures this, expressed using N3Logic with extensions. We state under KG some statements we assume to be present in the KG:

KG

NonProfit subClassOf org:Organization. AuditCo subClassOf org:Organization.
CertificationAgency subClassOf org:Organization. Donate subClassOf Action.
<ReliefOrg> type NonProfit. <AuditInc> type <CertificationAgency>.
<ReliefOrg> officer <George>. <AuditInc> officer <George>.

RuleEd1

{?O type org:Organization; hasOfficial: ?Ofc; hasFinancialRecord ?FR. ?Ofc type foaf:Person.
?FR assertions ?FRa. ?FRa log:semantics ?FRaS; ?FRAS log:includes { ?FR auditedBy ?Aud. ?Aud type AuditCo}.
?FR provenance ?PFr. ?PFr log:semantics ?PFrS. ?PFrS log:includes {prov:hasPrimarySouce ?DOCS}.
?DOCS assertions ?DOCSa. ?DOCSa log:semantics ?DOCSaS.
?DOCSaS log:includes {AuditCo certifiedBy ?CA. }. <TrustedGraphEd> author <Ed>
log:semantics ?TGEd. ?TGEd log:includes {?CA type CertificationAgency}}
=> {<TrustedGraphEd> :add {?O type:NonProfit}. }

This is a trust rule since it refers to KIs which are needed as input for Ed to take an action, "Donate".

Ed's second rule is in fact a business rule – he does not want to contribute to an organization in which George is an officer.

[6] https://www.w3.org/Submission/SWRL/

While this could be embedded in application code using the KG, we express it here as a rule as well, to facilitate discussion.

RuleEd2

{<TrustedGraphEd> author <Ed>; log:semantics ?TGEd. ?TGEd log:includes {?O type:NonProfit, officer:<George>}. <ruleEd1 author <Ed>}.
ruleEd1 assertions ?ARule1. ?Arule1 log:semantics ?Arule1S.?ARule1S log:includes {{ <act1> type Donate, recipient <?O>. <Ed> intends <act1>}
=> {<at> type Authorization, rule <RuleEd2>, action <act1>, value "Denied"}}

Next, we look at George's privacy rule. It is a privacy rule because it refers to an action over personal information about George.

RuleGeorge1

{G :is {?O type:NonProfit; officer:<George>. <TrustedGraphGeorge> author <George>; log:semantics ?TGG. ?TGG log:includes {?O officer:<George>. <ruleGeorge1> author <George>}.
ruleGeorge1 assertions ?AGeorge1. ?AGeorge1 log:semantics ?AGeorge1S. ?Ageorge1S log:includes { <act1> type Read; object ?G. ?A intends <act1>}
=> {<at> type Authorization; rule <RuleGeorge1>; action <act1>; value "Denied"}}

The stipulation expressed in the transparency legislation for non-profits to divulge its officers can be expressed as

KG

NonProfitAct type Law.

RuleTransp

{<ruleTransp> author <Congress>. <ruleTransp> assertions ?ATransp. ?ATransp log:semantics ?ATranspS. ?ATranspS log:includes{?G is {?O type Nonprofit; officer ?Ofi. <act1> type Read; object ?G. ?A intends <act1>}
=> {<at> type Authorization; rule <ruleTtransp>; action <act1>; value "Allowed"}}.
<ruleTransp> provenance ?PTransp. ?PTransp log:semantics ?PTranspS. ?PTranspS log:includes {<ruleTransp> prov:HasPrimarySource <NonProfitAct>}

We have also included in the provenance graph of RuleTransp a reference to the proper legal document, the text of the law (<NonProfitAct>) that is being interpreted by this rule.

In order to manage the conflict between RuleGeorge1 and RuleTransp, there is a meta-rule stipulating the latter has precedence over the former. This precedence relation is used in the Sort-by-Precedence function call in line 5 in Figure 1.

KG

PersonalPrivacyRule subClassOf Rule. Legislation subClassOf Rule.

MetaRule1

{<MetaRule1> assertions ?AMR1. ?AMR1 log:semantics ?AMR1S. ?AMR1S log:includes {{?R1 type PersonalPrivacyRule. ?R2 type Legislation} => {?R2 precedes ?R1}}

We can also define more precisely what is a PersonalPrivacyRule – it is any rule that uses a PersonalInformation property in its antecedent.

?RuleP assertions ?RulePA.?RulePA log:semantics ?RulePAS;?RulePAS log:includes {{?p1 ?r ?p2. (?p1 rdf:type Person OR ?p2 rdf:type Person)}}
=> {?r rdf:type PersonalInformationRelation}
{?RulePP assertions ?RulePAS. ?RulePPAS log:semantics ?RulePPASS;
?RulePPAS log:includes{?RulePP antecedent ?RulePPAA. ?RulePPAA log:semantics ?RulePPAAS ?RulePPAAS log:includes {{?p ?r ?q.(?r rdf:type PersonalInformationRelation)}
=>{?RulePP type PersonalPrivacyRule}}

The last scenario does not require any additional rules. It simply results in an "allowed" authorization, because no conflict arises between RuleGeorge1 and RuleTransp when applied to <AuditInc>, since it is not of type NonProfit. This illustrates a possible loophole in the legislation which could be avoided if the NonProfit legislation prohibited an organization from being audited by another organization having common officers.

6 Conclusions

We have presented a usage framework that explicates the various types of specifications that must be made to capture privacy, transparency and trust concerns. The framework also provides a better understanding of the relations between these concerns.

Trust entails determining which data items will be used to perform an action; privacy and transparency involve controlling who can perform an action over a data item. Trust is thus more fundamental, as privacy and transparency rules must be based on trusted data. Furthermore, privacy and transparency are seen as being different points along the control dimension, thus explaining the natural tension between the two.

We have shown how legal requirements, and other types of norms, which ultimately regulate the functioning of any application that uses the KG, can be incorporated into the KG itself.

The various nuances and interdependencies of these concerns we illustrated in a running example. One interesting point in the example is the fact that in spite of careful policies, loopholes in the regulations could allow undesired actions to take place.

As ongoing and future work, we are investigating implementation architectures to allow efficient and scalable usage control over existing KGs, as well as exploring the applicability to various domains.

7 ACKNOWLEDGMENTS

Daniel Schwabe was partially supported by a grant from CNPq.

REFERENCES

[1] Almendra, V. D. S., & Schwabe, D. (2006). Trust policies for semantic web repositories. In Proceedings of 2nd International Semantic Web Policy Workshop (SWPW'06), at the 5th International Semantic Web Conference (ISWC 2006) (pp. 17-31).

[2] Angella J. ;Kim, K; K.P. Johnson, Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook, Computers in Human Behavior, Volume 58, 2016, Pages 98-108, ISSN 0747-5632, http://dx.doi.org/10.1016/j.chb.2015.12.047.

[3] Artz, D; Gil, Y.; A survey of trust in computer science and the Semantic Web, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 5, Issue 2, 2007, Pages 58-71, ISSN 1570-8268, http://dx.doi.org/10.1016/j.websem.2007.03.002.(http://www.sciencedirect.com/science/article/pii/S1570826807000133)

[4] Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., Hendler, J.: N3Logic: A Logical Framework for the World Wide Web. Theory Pract. Log. Program. 8, 249–269 (2008).

[5] Bernstein , E. 2017. Making Transparency Transparent: The Evolution of Observation in Management Theory . Academy of Management Annals 11 (1): 217 – 66

[6] Bizer, C. ; Cyganiak, R.. Quality-Driven Information Filtering Using the WIQA Policy Framework. Web Semantics: Science, Services and Agents on the World Wide Web. 7. 1-10. 10.1016/j.websem.2008.02.005. (2009) - http://www.websemanticsjournal.org/index.php/ps/article/view/157/155

[7] Carlson, A., Betteridge, J., Wang, R.C., Hruschka Jr., E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 101–110 (2010)

[8] Castelfranchi, C., Falcone, R. Social Trust: A Cognitive Approach. In: Castelfranchi, C.; Yao-Hua Tan (Eds.): Trust and Deception in Virtual Societies. Springer-Verlag (2001).

[9] Castells, M., 2010. The Rise of the network society (2 ed.). Blackwell Publishers, Inc., Cambridge, MA, USA ISBN 978-1-405-19686-4. DOI:10.1002/9781444319514

[10] Davenport, T.H. and Beck, J.C., 2001. The attention economy: Understanding the new currency of business. Harvard Business Press.

[11] Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. SEMANTiCS,.CEUR 1695, available at http://ceur-ws.org/Vol-1695/paper4.pdf

[12] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečic, D. (2014): Introducing wikidata to the linked data web, Proc. ISWC 2014, Part I. LNCS, vol. 8796. Springer, Cham. https://doi.org/10.1007/978- 3- 319- 11964- 9

[13] Gerck, E.; Toward Real-World Models of Trust: Reliance on Received Information. Report MCWG-Jan22-1998, DOI: 10.13140/RG.2.1.2913.6727. http://www.safevote.com/papers/trustdef.htm.

[14] Grandison, T., & Sloman, M. (2000). A survey of trust in internet applications. IEEE Communications Surveys & Tutorials, 3(4), 2-16.

[15] Groth, Paul, Andrew Gibson, and Jan Velterop. "The anatomy of a nanopublication." Information Services & Use 30.1-2 (2010): 51-56.
[16] Heimstäd, M., Dobusch, L. 2018. Politics of Disclosure: Organizational Transparency as Multiactor Negotiation, Public Administration Review 78 (5), 727-738
[17] Kuhn, T., Meroño-Peñuela, A., Malic, A., Poelen, J.H., Hurlbert, A.H., Centeno, E., Furlong, L.I., Queralt-Rosinach, N., Chichester, C., Banda, J.M., Willighagen, E.L., Ehrhart, F., Evelo, C.T., Malas, T.B., & Dumontier, M. (2018). Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. CoRR, abs/1809.06532.availble at https://arxiv.org/pdf/1809.06532.pdf
[18] Laufer, C; Schwabe, D.; On Modeling Political Systems to Support the Trust Process, Proceedings of the Privacy Online 2017 Workshop (PrivON 2017), co-located with the 16th International Semantic Web Conference (ISWC 2017) Vienna, Austria, Out. 2017. CEUR-WS.org, 2017. v.1951. p.1 – 16
[19] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia-A large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web J. 6(2), 167–195 (2013)
[20] Lenat, D.B.: CYC: a large-scale investment in knowledge infrastructure. Commun. ACM 38(11), 33–38 (1995)
[21] Marwick, A., Lewis, R.. "Media Manipulation and Disinformation Online." Data & Society Research Institute, New York, 2017 - https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf
[22] Meijer, A. 2013. Understanding the Complex Dynamics of Transparency. Public Administration Review 73 (3): 429 – 439.
[23] Miller, G. A. 1995. WordNet: A lexical database for English. COMMUN ACM 38(11):39–41
[24] Paci, F., Squicciarini, A., Zannone, N.: Survey on Access Control for Community-Centered Collaborative Systems. ACM Comput. Surv. 51, 6:1–6:38 (2018)
[25] Pinyol, I.; Sabater-Mir, J.; Computational trust and reputation models for open multi-agent systems: a review. Artificial Intelligence Review 40:1–25 DOI 10.1007/s10462-011-9277-z (2013)
[26] Rifkin, J. (2011). The third industrial revolution: How lateral power is transforming energy, the economy, and the world. New York: Palgrave Macmillan.
[27] Schudson, M.;The Rise of the Right to Know: Politics and the Culture of Transparency, 1945-1973. Cambridge, MA: Harvard University Press, 2015.
[28] Schwab, K., 2017. The fourth industrial revolution. Crown Business.
[29] Schwabe, D., Laufer, C., Busson, A. Building Knowledge Graphs About Political Agents in the Age of Misinformation, http://arxiv.org/abs/1901.11408
[30] Sherchan, W., Nepal, S., and Paris, C. 2013. A Survey of trust in social networks. ACM Comput. Surv. 45, 4, Article 47 (August 2013), 33 pages. DOI: http://dx.doi.org/10.1145/2501654.2501661
[31] Smith, H.J., Dinev, T., Xu, H.: Information Privacy Research: An Interdisciplinary Review. MIS Quarterly. 35, 989–1015 (2011)
[32] Speer, R., Havasi, C.; "Representing General Relational Knowledge in ConceptNet 5, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), ELRA, 2012.
[33] Such, J.M., Criado, N.: Multiparty Privacy in Social Media. Commun. ACM. 61, 74–81 (2018)
[34] Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In: 16th International Conference on World Wide Web, pp. 697–706 (2007).
[35] Tavani, H.T.: Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy. Metaphilosophy. 38, 1–22 (2007).
[36] Tavani, H.T., Moor, J.H.: Privacy Protection, Control of Information, and Privacy-enhancing Technologies. SIGCAS Comput. Soc. 31, 6–11 (2001).
[37] United Nations, Universal Bill of Human Rights, Resolution A/RES/217(III)[A] http://unbisnet.un.org:8080/ipac20/ipac.jsp?session=14O243550E15G.60956&profile=voting&uri=full=3100023~!909326~!67
[38] Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base. Commun. ACM 57(10), 78–85 (2014)

An Attention-based Model for Joint Extraction of Entities and Relations with Implicit Entity Features

Yan Zhou
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
zhouyan@iie.ac.cn

Longtao Huang
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
huanglongtao@iie.ac.cn

Tao Guo
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
guotao@iie.ac.cn

Songlin Hu
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
husonglin@iie.ac.cn

Jizhong Han
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
hanjizhong@iie.ac.cn

ABSTRACT

Extracting entities and relations is critical to the understanding of massive text corpora. Recently, neural joint models have shown promising results for this task. However, the entity features are not effectively used in these joint models. In this paper, we propose an approach to utilize the implicit entity features in the joint model and show these features can facilitate the joint extraction task. Particularly, we use the hidden-layer vectors extracted from a pre-trained named entity recognition model as the entity features. Thus, our method does not need to design the entity features by hand and can benefit from the new development of named entity recognition task. In addition, we introduce an attention mechanism in our model which can select the informative parts of the input sentence to the prediction. We conduct a series of experiments on a public dataset and the results show the effectiveness of our model.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

Relation extraction; Implicit entity features; Attention mechanism; Neural network

Tao Guo is the corresponding author.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317704

ACM Reference Format:
Yan Zhou, Longtao Huang, Tao Guo, Songlin Hu, and Jizhong Han. 2019. An Attention-based Model for Joint Extraction of Entities and Relations with Implicit Entity Features. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3317704

1 INTRODUCTION

The extraction of entities and relations from unstructured text is an important task in information extraction and natural language processing (NLP). The goal of the task is to extract entities and their semantic relations from an unstructured input sentence. For example, in a sentence *"Donald Trump is the 45th and current president of the United States"*, *"Donald Trump"* and *"United States"* are the entities, the relation between them is *"president of country"*. The extracted information can be used for various kinds of downstream tasks such as question answering and knowledge base population. The methods for this task can be classified into two categories: pipelined models and joint models.

The pipelined models treat this problem as two separate tasks: named entity recognition (NER) [18, 19] and relation classification [22, 36]. These methods first identify the entity mentions in the sentence. And then, they take the identified mentions as correct entities and predict the relations between them. Therefore, the results of named entity recognition may have an impact on relation classification and lead to error propagation between them [14].

Different from the pipelined approaches, the joint models simultaneously identify the entity mentions and relations. These models always take an sentence as input and predict the entities and relations at the same time. Such joint methods have been proved to achieve better performance than the pipelined models. In earlier researches, the joint extraction systems typically rely on handcrafted

features, however the design of these features is time-consuming. With the capability of automatic feature learning, neural network (NN) based models are proposed to resolve the relation extraction problem [12, 17, 30, 37, 38]. These joint models can utilize the inter-dependence of the entities and relations. But existing studies generally employ word embeddings as input and do not exploit external entity features in their models. And we believe that the external entity features are beneficial to the joint extraction models, since the output triplets of the task are comprised by the entities and their relations.

In this paper, we propose an attention-based joint model enhanced with implicit entity features for the extraction task. In particular, we build our model upon the tagging scheme proposed by Zheng et al. [38]. Firstly, inspired by the work of Gao et al. [6], we dump the hidden layer vectors of a pretrained named entity recognition model as the entity features. These vectors contain rich entity information about each word, yet do not require manual design. We use these vectors as the entity features of our joint model. Through the utilization of these features, our method can easily benefit from the new development and the external training data of the NER task. In addition, when the model extracts the entities and relations, there are some tokens play a more important role than others. In the example sentence above, when the model predict the tag of word "Donald", the words representing the other entity (i.e. "United" and "States") and the word indicating the relations of the entities (i.e. "president") are more important than others. To utilize these information, we propose an attention mechanism in our model.

In summary, our main contributions are as follows:

- We propose an approach to integrate implicit entity features to the joint extraction task and show that these features can facilitate this task.
- We design an attention mechanism in our model, thus our model can focus on the informative words for the prediction.
- A series of experiments conducted on a public dataset demonstrate the effectiveness of our model.

The rest of this paper is structured as follows. We describe the related work of the relation extraction task in Section 2. We introduce the preliminary knowledge in Section 3. Section 4 describes our proposed model. The experiments and results are detailed in Section 5. In Section 6, we conclude this paper.

2 RELATED WORK

The task of relation extraction is to extract triplets that are composed of two entities and the relation between these two entities. There have been many studies on the extraction of entities and relations, and the methods can be roughly divied into two categories: the pipelined methods and the joint methods. The pipelined methods first perform the named entity recognition to identify entity mentions [18, 19], and then classify the relation between two entities [22, 36]. However, the pipelined methods ignore the dependency between two subtasks, which may lead to error delivery [14]. To solve this problem, the joint methods are proposed to extract entities and relations simultaneously. Therefore, the problem we focused is related to named entity recognition, relation classification and joint extraction of entities and relations. Our method is also related to attention models.

2.1 Named Entity Recognition

Named entity recognition (NER) has a long history in the field of natural language processing. Earlier researches focused on linear statistical models, such as Hidden Markov Models (HMM) [39] and Conditional Random Fields (CRF) [33]. These models relied on hand-crafted features, gazetteers and other external resources to perform well. Recently, sereval neural models have been proposed for NER . Collobert et al. [5] used a CNN over a sequence of word embeddings with a CRF layer on top. Huang et al. [11] replaced the CNN encoder in Collobert et al. with bidirectional LSTM encoder. Lample et al. [13] and Chiu & Nichols [4] introduced neural models with additional bidirectional LSTM and CNN encoders to encode character-level features. Marek Rei et al. [20] presented an attention-based neural model to improve the performance by the character-level extentions. To reduce the amount of labeled training data in NER, Yanyao Shen et al. [25] combined deep learning with active learning.

2.2 Relation Classification

Relation classification is a common task in natural language processing. Apart from a few unsupervised clustering methods [3, 8], the majority of studies on relation classification have been supervised. Zeng et al. [34] proposed a convolutional deep neural network to extract lexical and sentence level features. And then, these features were fed into a softmax classifier to predict the relations. Zhang and Wang [35] investigated a temporal structured RNN with only words as input. Zhou et al. [40] used bidirectional LSTM with attention mechanism to capture the most important semantic information in a sentence. Wang et al. [29] proposed a novel convolutional neural network architecture for relation classification, relying on two levels of attention in order to better discern patterns in heterogeneous contexts.

2.3 Joint Extraction of Entities and Relations

There are a number of researches on the joint extraction of entities and relations. Most of the joint methods are feature-based models [14, 21], in which the design of handcrafted features is time-consuming. Recently, deep learning methods provide an effective way to reduce the manual work. For example, an relation extraction model depending on both word sequence and dependency tree structure was proposed to extract entities and relations [17]. Katiyar and Cardie [12] presented a novel recurrent neural network to extract semantic relations between entity mentions without having access to dependency trees. Though their proposed methods used a single model, the model needs to identify entities first and then extracts the relations between them. In order to better utilize the interactions between the outputs, some novel schemes that can jointly decode the entities and relations were proposed. Zheng et al. [38] introduced a new tagging scheme, based on what the task of entities and relations extraction is transformed into the sequence tagging problem. And they proposed an end-to-end model to solve

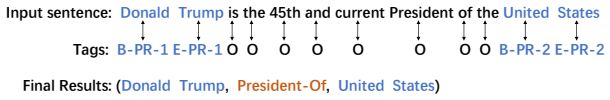


Figure 1: Gold standard annotation for the example sentence, where "PR" is short for "President-Of".

the problem. Wang et al. [30] converted the joint task into a directed graph by designing a novel graph scheme and proposed a transition-based approach to achieve joint learning through joint decoding. Though these joint models show promising results for the extraction task, they do not take advantage of external entity features.

2.4 Attention Models in NLP

Attention-based neural networks are proposed to obtain a representation weighted by the importance of tokens in the sequence. Such models have been successfully applied on many NLP applications. Bahdanau et al. [1] proposed an encoder-decoder framework with attention mechanism for machine translation. Wang et al. [31] presented an attention-based neural network for aspect-level sentiment classification. The attention mechanism can concentrate on different parts of a sentence when different aspects were taken as input. Rush et al. [23] utilized a local attention-based model that generated each word of the summary conditioned on the input sentence for sentence summarization. Yang et al. [32] introduced a hierarchical attention network for document classification. Seo et al. [24] introduced the Bi-Directional Attention Flow network for machine comprehension, which represented the context at different levels of granularity and used bidirectional attention flow mechanism to obtain a query-aware context representation.

3 PRELIMINARY KNOWLEDGE

In this section, we introduce the preliminary knowledge related to our proposed approach. We first describe the task of relation extraction in Section 3.1. Then, we briefly introduce the tagging scheme in Section 3.2. Finally, we provide a brief description of LSTM in Section 3.3.

3.1 Task Definition

The goal of relation extraction is to mine entities and their relations from unstructured texts. As shown in Figure 1, the input of the task is a sentence *"Donald Trump is the 45th and current President of the United States"* and the output of can be represented as a triplet $(Donald\ Trump, President-Of, United\ States)$. The triplet is comprised of two entities and their relation. For the above example, *"Donald Trump"* and *"United States"* are the entities, and *"President-Of"* denotes the relation between the them.

3.2 Tagging Scheme

The tagging scheme is proposed by Zheng et al. [38] to extract the entities and their relations simultaneously. Based on this tagging scheme, the information between the output entities and relations can be fully exploited. Figure 1 presents an example of the tagging scheme. Tag 'O' represents that the corresponding word is independent of extracted results. The other tags which represent the elements in the triplet are named as relational tag. The relational tag is comprised of three parts: the word position in entity, the relation type and the relation role. The position of a word in entity is represented in "BIES" (Begin, Inside, End, Single) format. The relation type is obtained from a set of predefined relations. The relation role is represented by '1' or '2' which means the first or the second entity in the relation, respectively. According to the tag sequence, we can directly get the entities with its relation in the form of a triplet $(Entity_1, RelationType, Entity_2)$. For the sentence presented in Figure 1, we can extract the relation triplet $(Donald\ Trump, President-Of, United\ States)$. Then, based on this tagging scheme the goal of relation extraction is to predict the tag of each token.

3.3 LSTM

Recurrent neural networks (RNNs) are a family of neural networks which are capable of processing sequential input of unbounded and arbitrary length. Long Short Term Memory Networks (LSTMs) [9] are a special kind of RNNs. They take a sequence of vectors $(x_1, x_2, \ldots, x_n)$ as input and return another sequence $(h_1, h_2, \ldots, h_n)$ that represents some information about the sequence at every time step. LSTMs incorporate gating functions at each time step to allow the network to forget, remember and update contextual memory and mitigate problems like vanishing gradient. We use the following implementation:

$$i_t = \sigma(W_{xi}x_t + b_{xi} + W_{hi}h_{t-1} + b_{hi}) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + b_{xf} + W_{hf}h_{t-1} + b_{hf}) \quad (2)$$
$$g_t = tanh(W_{xg}x_t + b_{xg} + W_{hg}h_{t-1} + b_{hg}) \quad (3)$$
$$o_t = \sigma(W_{xo}x_t + b_{xo} + W_{ho}h_{t-1} + b_{ho}) \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$
$$h_t = o_t \odot tanh(c_t) \quad (6)$$

where σ is the sigmoid function, and $\odot$ is the element-wise product.

4 METHODOLOGY

In this section, we describe the proposed approach in detail. Following recent studies, we extract the entities and relations jointly in a single model. In particular, we follow zheng et al. [38], casting the joint extraction task as a sequence tagging problem. Given an input sentence $S = (w_1, w_2, \ldots, w_n)$, the output of the model is a sequence of tags $T = (t_1, t_2, \ldots, t_n)$ corresponding to the sentence. Based on these tags, we can directly get the entities and relations. Figure 2 shows the overall architecture of our model which has four main components: an embedding layer, an encoding layer, an attention layer and a decoding layer. The embedding layer represent each input token by multi-level features. The encoding layer learns the contextual information of the input tokens. The attention layer selects the important information from the output of the encoding layer. The decoding layer predicts the tag for each word based on the vector computed by the attention layer.

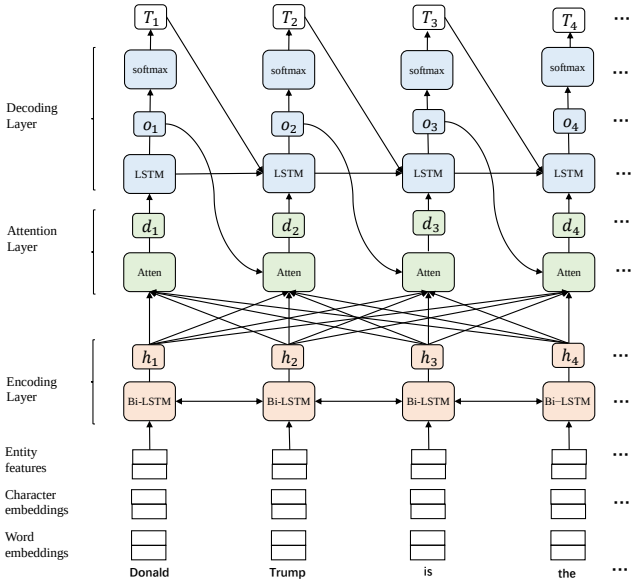


Figure 2: Our network structure for joint extraction of entities and relations.

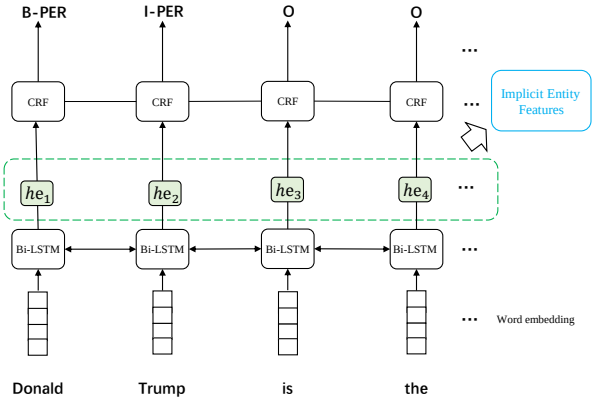


Figure 3: Entity Features Generation Model.

4.1 Features

We utilize the word embeddings, character embeddings and implicit entity features to represent the input words.

Word Embeddings. For each word w_t in S, we look up the embedding matrix to get its word-level representation ww_t. It has been reported that word embeddings learned from significant amounts of unlabeled data are far more satisfactory than the randomly initialized embeddings [5]. Therefore, we initialize the matrix by word embeddings pretrained by word2vec model [15].

Character Embeddings. It has been demonstrated that character embeddings are effective to handle the out-of-vocabulary problem for several NLP tasks such as named entity recognition [20] and dependency parsing [2]. Thus, we use the character sequence of each input token to produce its character-level representation. Each word is broken up into individual characters, and these characters are mapped to a sequence of character embeddings $(c_1, c_2, \ldots, c_L)$. After that, we adopt a bidirectional LSTM to generates the character embedding wc_t for the word w_t.

Implicit Entity Features. We introduce a method to integrate implicit entity features for the joint extraction of entities and relations in this section. Intuitively, the external entity features are useful for the joint extraction task. However, the previous neural joint methods do not consider this kind of features. Since designing these features manually is always time consuming, we propose to incorprate the automatically generated entity features in our model. Inspired by the work of Gao et al. [6], we leverage the hidden vectors in a pre-trained named entity recognition model to generate the implicit entity features. We first pre-train a named entity recognition model on an existing named entity recognition dataset. And then, the input sentences are fed into this model, and the hidden vectors which contain rich entity information are dumped as the entity features. Thus, our method dose not need to manually design these features and can easily benefit from the latest developments of NER task.

As shown in Figure 3, we model named entity recognition as a sequence tagging problem and employ an simplified BLSTM-CRF model as our entity features generation model. First of all, each token is represented by its word embedding. Then we feed these vectors into a bidirectional LSTM. Bidirectional LSTM contains two separate LSTMs to capture both the past and future information, both of the two information are beneficial to the named entity recognition. For the task of named entity recognition, it is important to consider the constrains between tags in neighborhood (e.g., I-PER can not follow B-LOC). And Conditional random fields (CRF) are effective to learn these constrains. Therefore, a standard CRF layer is used on the top of the model to jointly predict the tags.

To obtain the implicit entity features, we pre-train the above model on existing named entity recognition datasets first. And then, for a given sentence $(w_1, w_2, \ldots, w_n)$, we directly dump a sequence of hidden states $(he_1, he_2, \ldots, he_n)$ from the bidirectional LSTM layer of the model:

$$(he_1, he_2, \ldots, he_n) = BLSTM_{entity}(w_1, w_2, \ldots, w_n) \quad (7)$$

In this way, we get the implicit entity features representation he_t of the t-th word in the input sentence.

4.2 Encoding Layer

In the encoding layer, we adopt a bidirectional LSTM which can combine the forward and backward context of a word to encode the sentence. We use x_t to denote concatenation of ww_t, wc_t and he_t. Thus, the input of the encoding layer is a sequence of vectors $(x_1, x_2, \ldots, x_n)$. And this layer receives these vectors as input and computes the t step hidden state h_t as follows:

$$\overrightarrow{h_t} = LSTM_{encoder}(x_t, \overrightarrow{h_{t-1}}) \quad (8)$$

$$\overleftarrow{h_t} = LSTM_{encoder}(x_t, \overleftarrow{h_{t+1}}) \quad (9)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (10)$$

Thus, we obtain the representation of each input token containing the contextual information in the encoding layer.

4.3 Attention Layer

In order to effectively utilize the information of the relevant tokens in the sentence, we design an attention layer as depicted in Figure 4. First, we propose a Tag-Aware attention for the purpose of selecting the relevant information of the word. Then, we adopt a gate to fuse this relevant semantic information and the word information obtained from the encoding layer. The details of this layer will be illustrated as follows.

Tag-Aware Attention The key components for predicting the tag of each input word may appear anywhere in the sentence. To better utilize these informative words, we employ an attention mechanism which is named as Tag-Aware Attention. The Tag-aware attention allows the model to select the relevant parts of the source sentence for the prediction of the tag.

In order to compute the relevant word representation of the t-th word, we first calculate an attention score $\alpha_{j,t}$ to the j-th word $(1 \leq j \leq n; j \neq t)$ as:

$$e_{j,t} = h_j \cdot o_{t-1} \tag{11}$$

$$\alpha_{j,t} = \frac{exp(e_{j,t})}{\sum_{k=1}^{t-1} exp(e_{k,t}) + \sum_{k=t+1}^{n} exp(e_{k,t})} \tag{12}$$

where h_j is the hidden state of the j-th word calculated by the encoding layer, and o_{t-1} represents the output status computed by the decoding layer. After that the vector ha_t representing the informative words in the sentence is computed as the summation vector weighted by $\alpha_{j,t}$:

$$ha_t = \sum_{j=1}^{t-1} \alpha_{j,t} h_j + \sum_{j=t+1}^{n} \alpha_{j,t} h_j \tag{13}$$

Fusion Gate Based on the attention vector ha_t and the word vector obtained from the encoding layer h_t, we design a fusion gate to combine them. When predicting the tag of a word, the gate allows the model to trade off the information used from ha_t and h_t, both of which are important to the successful prediction.

Therefore, instead of just concatenating ha_t with the vector h_t, the two vectors are added together using a weighted sum. And the weights are predicted by a two-layer network. Let d_t denotes the output of the attention layer corresponding to the t-th word in the sentence, we calculate it as follows:

$$g_t = \sigma(W_g^{(3)} tanh(W_g^{(1)} ha_t + W_g^{(2)} h_t)) \tag{14}$$

$$d_t = g_t ha_t + (1 - g_t) h_t \tag{15}$$

where $W_g^{(1)}, W_g^{(2)}, W_g^{(3)}$ are weight matrices for calculating g_t, and σ is the logistic function. The vector g_t acts as the weight between ha_t and h_t, which makes the model can dynamically merge the two vectors. In addition, g_t is computed for every word in the sentence, which allows the model to be more flexibly for making different decisions for diverse words.

4.4 Decoding Layer

In the decoding layer, both the tag embeddings and the vectors computed by the attention layer are used as the input. And we adopt LSTM to learn the dependency of the tags and generate

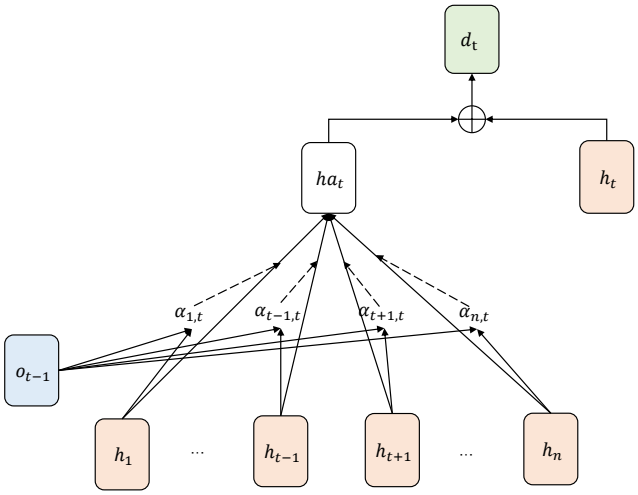


Figure 4: Overview of the Attention Mechanism.

vectors representing the output states:

$$\overrightarrow{o_t} = LSTM_{decoder}(d_t + W_T T_{t-1}, \overrightarrow{o_{t-1}}) \tag{16}$$

where o_t represents the hidden state of the t step, d_t denotes the vector from the attention layer; T_{t-1} is the tag embedding directly converted by the *t-1*-th tag, W_T is the weight matrix. Finally we adopt a softmax classifier to compute normalized entity tag probabilities based on the vector o_t:

$$s_t = W_s o_t + b_s \tag{17}$$

$$p(tag_t|S, tag_{t-1}) = \frac{exp\ (s_t)_{tag_t}}{\sum_{tag'_t \in TAG} exp\ (s_t)_{tag'_t}} \tag{18}$$

where W_s denotes the weight matrix, b_s is the bias term, tag_t represents the tag of the t-th word and TAG is the tag set.

4.5 Objective Function

We use the bias training goal presented in [38] to maximize the log-likelihood of the data:

$$L = max \sum_{m=1}^{|\mathbb{D}|} \sum_{t=1}^{L_m} (log(p(tag_t^m|S_m, tag_{t-1}^m)) \cdot I(O) + \alpha \cdot log(p(tag_t^m|S_m, tag_{t-1}^m)) \cdot (1 - I(O))) \tag{19}$$

where $|\mathbb{D}|$ is the size of training set, L_m is the length of sentence S_m, tag_t^m denotes the tag of the t-th word in sentence S_m, $I(O)$ is a switching function to distinguish the loss of tag 'O' and relational tags:

$$I(O) = \begin{cases} 1, & if\ tag = 'O' \\ 0, & if\ tag \neq 'O' \end{cases} \tag{20}$$

α is the bias weight which is used to control the influence of relational tags.

Table 1: Dimension sizes

Model	Network Structure	Size
Our Model	Word Embedding	300
	Char Embedding	20
	$LSTM_{char}$	25
	Entity Features	100
	$LSTM_{encoder}$	300
	$LSTM_{decoder}$	600
	Tag Embedding	50
Entity Features Generation Model	Word Embedding	300
	$LSTM_{entity}$	50

Table 2: Comparison with baselines

Method	**Prec.**	**Rec.**	**F1**
FCM	0.553	0.154	0.240
DS+logistic	0.258	0.393	0.311
LINE	0.335	0.329	0.332
MultiR	0.338	0.327	0.333
DS-Joint	0.574	0.256	0.354
CoType	0.423	0.511	0.463
LSTM-LSTM-Bias	0.615	0.414	0.495
Transition-Based	0.643	0.421	0.509
Our model	0.640	0.464	**0.538**

5 EXPERIMENTS

5.1 Experiment Setup

Dataset and Evaluation Metrics. We evaluate our proposed model with the public dataset NYT[1] which is developed by Ren et al. [21]. There are 353k triplets in the training data and 3880 triplets in the test set. The test set is manually annotated to guarantee its quality. The size of the relation type in the dataset is 24. In addition, we use CoNLL-2003 dataset [28] to pre-train the entity features generation model.

We adopt Precision(Prec), Recall(Rec) and F1-scores to evaluate our method. E1, E2 and R represent the first entity, the second entity and the relation type in the triplet, respectively. A triplet is regarded as correct when E1, E2 and R are all correct [21, 30, 38]. Similar to the previous work, we randomly sample 10% data from the test set as our validation set and use the remaining data as evaluation.

Hyperparameters. For experiments, we initialize our word embeddings with 300-dimensional embedding vectors trained on NYT corpus. For the tag embedding, we initialize them by random weights. We update the parameters of our model by backpropagation using RMSprop [27] with learning rate 5×10^{-4} and mini-batch size 50. For regularization, we employ dropout operation with dropout rate of 0.3 for the word embeddings. For the entity features generation model, we select RMSprop with learning rate 0.01 to train the model. The dimension of the vectors in our model and the entity features generation model are shown in Table 1. The bias term is set to 10.

5.2 Comparison with Baselines

Baselines. We compare our model against several baseline models for relation extraction, which can be divided into the following categories: the pipelined methods, the joint extraction methods and the end-to-end methods.

For the pipelined methods, the NER results are obtained by CoType [21], and then several classical relation classification methods are applied to detect the relations. The pipelined methods compared in this paper are as follows:

- **FCM** [7] a method which combines unlexicalized hand-crafted features with learned word embeddings.
- **DS+logistic** [16] combines distant supervision and syntactic parse features to improve the performance of relation extraction.
- **LINE** [26] is a network embedding method, which is suitable for arbitrary types of information networks.

We compare our model with the following joint extraction methods:

- **MultiR** [10] presents a novel approach for multi-instance learning with overlapping relations.
- **DS-Joint** [14] presents an incremental joint framework to simultaneously extract entity mentions and relations using structured perceptron with efficient beam-search.
- **CoType** [21] formulates the joint entity and relation mention typing problem as a global embedding problem.

The end-to-end methods compared with our model are as follows:

- **LSTM-LSTM-Bias** [38] converts the entities and relations extraction to a sequence tagging problem and jointly extracts them by an end-to-end model.
- **Transition-Based** [30] converts the joint task into a directed graph by designing a novel graph scheme and proposes a transition-based approach to extract the relations.

Table 2 presents the performance of our model as well as all of the baselines mentioned above. We observe that our end-to-end model achieves significant improvements over all the baselines in F1 score. In particular, it achieves 2.9% improvement over the best end-to-end methods. This demonstrates the effectiveness of our model on the extraction of entities and relations. From the above table, we can find that the joint methods perform better than the pipelined methods. This is because the joint models can avoid error delivery problem. The end-to-end models outperform the joint models. This indicates the joint decoding which can take advantage of the dependencies of the output is important for the extraction task. Our model achieves better performance compared with other end-to-end models. One reason is that employing entity features as the complementary to the standard word embeddings, our model can take advantage of extra entity information. Meanwhile, our model is able to focus on the informative parts of the input through the applying of Tag-Aware attention.

[1] https://github.com/shanzhenren/CoType/tree/master/data/source/NYT

Table 3: Ablation results

Method	Prec.	Rec.	F1
Our model	0.640	0.464	**0.538**
-Implicit Entity features	0.604	0.456	0.521
-Attention Layer	0.601	0.450	0.515
-Fusion Gate+Concat	0.599	0.469	0.526

5.3 Ablation Tests

To show the effectiveness of each components proposed in our model, we conduct a set ablation experiments and the results are presented in Table 3. The value of F1 drop 1.7%, when we do not use the implicit entity features in our model. This result shows that external entity features which are not utilized in other joint models are useful for the joint extraction task. The performance of the model will drop when the attention layer is removed. This demonstrates that the informative words in the sentence play an important role for the correct prediction. The last line presents the results, when we replace the fusion gate by merely concatenating the related word vector to the vector computed by the encoding layer. And the results prove that incorporating the relevant information of the sentence in a reasonable manner can further improve the performance.

To give a detailed analysis of the effect of the components, we show the results on different elements of the extraction triplet. E1, E2 and R represent the first entity, the second entity and the relation type in the triplet, respectively. If both the head offsets of E1 and the relation type R are correct, then the instance of (E1, R) is correct. Similarly, if both the head offsets of E2 and the relation type R are correct, then the instance of (E2, R) is correct. Regardless of relation type R, if both the head offsets of two corresponding entities are correct, the instance of (E1, E2) is correct.

Table 4, shows the results on different triplet elements. First of all, when we remove one of the component from the model, the performance of the three types triplet elements will decline. This indicates the proposed components can enhance both the entity recognition and the relation classification. Compared with the results on one of the the entities is not considered (i.e. (E1, R) and (E2, R)), the attention layer has a greater impact than the implicit entity features. This is because the informative words in the context is more important for the prediction of the relation. And from the last line of the table, we find that the effect of implicit entity features is much more than the attention layer when merely evaluating the entities of the triplet.

5.4 Impact of Implicit Entity Features Size

We make a detailed evaluation on the influence of implicit entity features size. Figure 5 shows the performance of our model when we change the dimension of the implicit entity features. We use the size of entity features in the following set {50, 100, 150, 200, 250}. From the figure we find that our model performs best when the size of entity features vector is set to 100. We can also see that the value of F1 is relatively stable when varying the dimension of implicit entity features. This demonstrates the robustness of the implicit entity features to our model.

Table 4: Ablation results on triplet elements

Elements	Method	Prec.	Rec.	F1
(E1, R)	Our Model	0.684	0.515	**0.588**
	-Implicit Entity Features	0.670	0.509	0.579
	-Attention Layer	0.669	0.501	0.573
(E2, R)	Our Model	0.663	0.499	**0.569**
	-Implicit Entity Features	0.642	0.488	0.555
	-Attention Layer	0.633	0.475	0.543
(E1, E2)	Our Model	0.667	0.501	**0.572**
	-Implicit Entity Features	0.646	0.490	0.558
	-Attention Layer	0.658	0.493	0.564

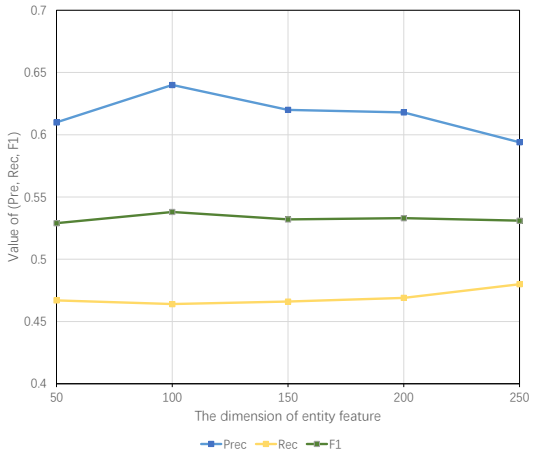


Figure 5: The results predicted by our model on different dimension of the entity feature.

5.5 Attention Visualization

We give a case study to illustrate the power of our attention mechanism. In Figure 6, we visualize the attention weights when our model predicts relational tags. We find that our model focuses more on the informative words (e.g. words stand for the entities in the triplet and words indicate the relation between the entities), when it predicts relational tags. For example, when predicting the tag of "Chad", our model pays more attention on the words "Hurley, executive, YouTube, company" which are essential for the tag "B-PC-1". Thus, our model can effectively utilize the informative words in the input sentence.

6 CONCLUSION

In this paper, we propose an attention-based model enhanced with implicit entity features for the joint extraction of entities and relations. We propose an approach to leverage the implicit entity

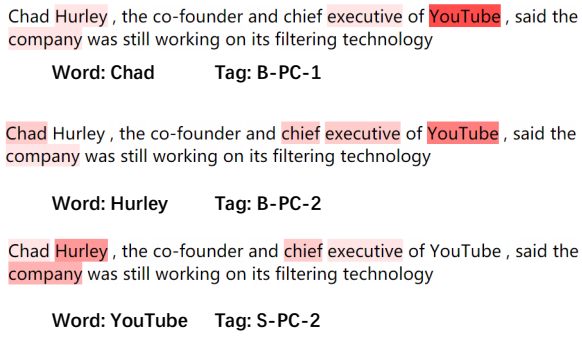


Figure 6: Visualization of attention weights when our model predicts the tags of the words {"Chad", "Hurley", "YouTube"}. We can extract the relation triplet $(Chad\ Hurley, Person-Company, YouTube)$ **from the example sentence. "PC" is short for "Person-Company" and darker color means higher weight.**

features in the joint extraction model. Thus, our model can take advantage of the entity features and dose not need to manually design them. In addition, we design a Tag-Aware attention mechanism which enables our model to select the informative words to the prediction. And our final model achieves competitive performance on a public dataset.

ACKNOWLEDGEMENTS

This research is supported in part by the National Key Research and Development Program of China (No. 2017YFB1010000) and the National Natural Science Foundation of China (No. 61702500).

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* (2015).
[2] Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2015. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 349–359.
[3] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2005. Unsupervised feature selection for relation extraction. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.
[4] Jason PC Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4 (2016), 357–370.
[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, Aug (2011), 2493–2537.
[6] Yuze Gao, Yue Zhang, and Tong Xiao. 2017. Implicit Syntactic Features for Target-dependent Sentiment Analysis. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 516–524.
[7] Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved Relation Extraction with Feature-Rich Compositional Embedding Models. In *EMNLP*. 1774–1784.
[8] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 415.
[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[10] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*. 541–550.
[11] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).
[12] Arzoo Katiyar and Claire Cardie. 2017. Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees. In *ACL*. 917–928.
[13] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT*. 260–270.
[14] Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 402–412.
[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[16] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*. 1003–1011.
[17] Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *ACL*. 1105–1116.
[18] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.
[19] Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 147–155.
[20] Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016. Attending to Characters in Neural Sequence Labeling Models. In *COLING*. 309–318.
[21] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. CoType: Joint extraction of typed entities and relations with knowledge bases. In *WWW*. 1015–1024.
[22] Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 256–259.
[23] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 379–389.
[24] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *ICLR*.
[25] Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. 252–256.
[26] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.
[27] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.
[28] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 142–147.
[29] Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. (2016).
[30] Shaolei Wang, Yue Zhang, Wanxiang Che, and Ting Liu. 2018. Joint Extraction of Entities and Relations Based on a Novel Graph Scheme. In *IJCAI*.
[31] Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.
[32] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.
[33] Lin Yao, Chengjie Sun, Shaofeng Li, Xiaolong Wang, and Xuan Wang. 2009. CRF-based active learning for Chinese named entity recognition. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*. IEEE, 1557–1561.
[34] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2335–2344.
[35] Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006* (2015).
[36] Zhu Zhang. 2004. Weakly-supervised relation classification for information extraction. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 581–588.
[37] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* 257 (2017), 59–66.
[38] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. In *ACL*. 1227–1236.

[39] GuoDong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. Association for Computational Linguistics, 473–480.
[40] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 207–212.

LA-Web 2019 Chairs' Welcome

It is our great pleasure to welcome you to The Latin American Web Congress (LA-WEB) co-located with THE WEB CONFERENCE which will take place in San Francisco, CA. USA, on May 14, 2019.

LA-WEB was born in 2003 as an international scientific venue that aims towards gathering the community of Web researchers and practitioners worldwide, working on topics of importance to the Latin American region. In addition, this year we have extended LA-WEB's scope to also include Web-centered research from Latin American institutions and from Latin American authors around the world. We expect that the combination of LA-WEB and The Web Conference will create a synergetic atmosphere between the Latin American Web community and the global community of leading Web researchers.

LA-WEB was designed to host both mature work (full papers) and work that is at early stages, which can benefit from discussion among peers (short papers). In particular, for this edition, after a careful review process, 6 full papers, 8 short papers, and 2 presentation-only-posters, were selected to be included in the technical program. The program also included keynote talks by two renowned Latin American researchers in their fields, namely Carlos Castillo and Jussara Almeida, who will cover exciting topics related to Web technologies and their impact in Latin America. To further foster discussion among the conference attendees, we will also host a poster session with all accepted papers. For closing the event there will be a final discussion session for all participants to address open problems and challenges.

We would like to kindly thank, in no particular order, all of the authors that submitted their work to LA-WEB, our outstanding Program Committee, and also The Web Conference organization, in particular, the Workshop Chairs and General Chairs. We are also grateful to Demi Getschko, Vagner Diniz, Diogo Cortiz and Tanara Lauschner, who were the first to raise the idea of hosting a co-located event with TWC 2019 dedicated to the Latin American community. As if that were not enough they have also been extremely helpful by providing their support in terms of helping secure funding and also with logistics. We also extend a special thank you to Ricardo Baeza-Yates, for his constant encouragement and advice in the task of creating an international venue for the Latin American Web community.

Furthermore, this event would have not been possible without the generous support of our sponsors, who have made it possible for all authors of accepted papers to travel and attend LA-WEB and The Web Conference 2019: NIC.br, CGI.br, NIC.cl, SCCC, ISOC Brasil, and IMFD.

We are looking forward to future editions of LA-WEB in hopes that it will establish itself in time as a high-quality venue for Latin American Web research from all over the world.

San Francisco, May 2019

Barbara Poblete
Program Committee Co-Chair
Universidad de Chile

Altigran S. da Silva
Program Committee Co-Chair
Universidade Federal do Amazonas

LA-Web 2019 Organization

THE WEB CONFERENCE

Program Committee Co-Chairs:
Barbara Poblete *(Universidad de Chile, Chile)*
Altigran S. da Silva *(Universidade Federal do Amazonas, Brazil)*

Organization Committee:
Diego Cortiz *(NIC.br, Brazil)*
Tanara Lauschner *(Universidade Federal do Amazonas, Brazil)*
Vagner Diniz *(NIC.br, Brazil)*

Program Committee:
Alfredo Sánchez, LANIA, Mexico
Andreas Kaltenbrunner, NTENT, Spain
Benjamin Piwowarski, CNRS/LIP6, France
Claudia López, USM, Chile
Claudio Gutierrez, UChile, Chile
Debora Donato, Mix Tech, USA
Denis Parra, PUC, Chile
Diego Arroyuelo, USM, Chile
Diogo Cortiz, PUC/SP, Brazil
Edgar Chávez, CICESE, Mexico
Eduardo Graells-Garrido, UDD, Chile
Fabrizio Silvestri, Facebook, UK
Felipe Bravo-Marquez, U Waikato, NZ
Francisco Gutierrez, UChile, Chile
Gonzalo Navarro, UChile, Chile
Gustavo Rossi , UNLP, Argentina
Jorge Perez, UChile, Chile
Juan Sequeda, Capsenta, USA
Jussara M. Almeida, UFMG, Brazil
Karin Breitman, Rio Tinto, Brazil
Luciano Barbosa, UFPE, Brazil
Luis Olsina, UNLPAM, Argentina
Maria Esther Vidal, USB, Venezuela
Marcelo Mendoza, USM, Chile
Marcos Gonçalves, UFMG, Brazil
Mirella M. Moro, UFMG, Brazil
Miriam Fernandez, Open University, UK
M-Dyaa Albakour, Signal Media, UK
Omar Alonso, Microsoft, USA
Pablo Castells, UAM, Spain
Rossano Schifanella, U Turin, Italy
Ruth Garcia-Gavilanes, Skyscanner, UK
Sole Pera, Boise State University, USA
Tanara Lauschner, UFAM, Brazil
Vanessa Peña-Araya, Inria Saclay, France
Wagner Meira Jr, UFMG, Brazil

Misinformation Dissemination on the Web

Jussara M. Almeida
Computer Science Department, Universidade Federal de Minas Gerais, Brazil
jussara@dcc.ufmg.br

ABSTRACT

Misinformation dissemination is a topic that has gained a lot of attention from academia and public media, in general. Despite a rich literature on strategies to detect and mitigate this phenomenon, the problem still persists with impact on several sectors of the society. In this talk, I will discuss the problem, revise existing approaches as well as discuss challenges to properly address it. I will also discuss recent results of our group on the investigation of misinformation spread on WhatsApp.

CCS CONCEPTS

• **Information systems → Collaborative and social computing systems and tools**; **Social networks**; *Data analytics.*

KEYWORDS

misinformation dissemination, online social networks, automatic detection

ACM Reference Format:
Jussara M. Almeida. 2019. Misinformation Dissemination on the Web. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316450

1 OUTLINE

The great popularity of various Web applications, notably social media and online social networks, has greatly contributed to turn it into one of the major sources of information for a large portion of the population. The easiness at which one can search, find and spread a piece of information, in various media types, has unquestionably revolutionized the processes guiding information dissemination in modern society. A side effect is that an old "phenomenon", gossiping and rumors, or more broadly speaking, misinformation dissemination, has gained scale and speed.

Misinformation, a broad term often used to refer to false or inaccurate information possibly spread with intention to deceive, is a topic that has attracted a lot of attention in the recent years [1–4, 11]. There is a rich literature addressing this topic in various applications and domains, offering analyses, characterizations, and potential countermeasures (e.g., [5–10, 12, 13], to name a few examples). Yet, the problem persists as user behavior evolves and new applications emerge.

In this talk, I will discuss misinformation on the Web, showing examples and practical implications, and related phenomena that contribute to its spread. I will revise some existing solutions to detect misinformation and their disseminator, and discuss challenges to properly address the problem, covering aspects related to the role of user participation and characteristics of emerging applications that may facilitate their spread.

Finally, focusing specifically on WhatsApp, a messaging app of enormous popularity in various countries, I will present some recent results on the spread of misinformation on the platform and how it differs from the other contents, hinting at possible approaches to mitigate the process.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316450

ACKNOWLEDGMENTS

The author thanks CNPq, CAPES and FAPEMIG for partial financial support.

REFERENCES

[1] H. Allcott and M. Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] A. Bessi and E. Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).
[3] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22, 8 (2017).
[4] Adam Fourney, Miklos Z. Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proc. of the Conference on Information and Knowledge Management (CIKM).*
[5] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, and others. 2018. Fake news vs satire: A dataset and analysis. In *Proc. of the Conference on Web Science (WebSci).*
[6] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor Has It: Identifying Misinformation in Microblogs. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMLP).*
[7] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019).
[8] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proc. of the Web Conference (WWW).*
[9] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[10] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it Hoax: Automated fake news detection in social networks. In *Proc. of the Workshop on Data Science for Social Good (SoGood).*
[11] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[12] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proc. of the 11th ACM International Conference on Web Search and Data Mining (WSDM '18).* ACM, New York, NY, USA, 637–645. DOI: http://dx.doi.org/10.1145/3159652.3159677
[13] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Emotional dynamics in the age of misinformation. *PloS one* 10, 9 (2015), e0138740.

Algorithmic Bias in Rankings

Carlos Castillo
chato@acm.org
Universitat Pompeu Fabra
Barcelona, Catalonia, Spain

ABSTRACT

In this talk we refer to bias in its everyday sense, as a prejudice against a person or a group, and ask whether an algorithm, particularly a ranking algorithm, can be biased. We begin by defining under which conditions this can happen. Next, we describe key results from research on algorithmic fairness, much of which studies automatic classification by a supervised learning method. Finally, we attempt to map these concepts to rankings and to introduce new, ranking-specific ways of looking at algorithmic bias.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Data mining*.

KEYWORDS

bias, discrimination, ranking

ACM Reference Format:
Carlos Castillo. 2019. Algorithmic Bias in Rankings. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316451

1 OUTLINE

Concerns about biased (in the sense of prejudiced) algorithms have been motivated by research that has exposed "racist" [1] and "sexist" [3, 8] algorithms. In this talk, we will start by defining specifically under which circumstances an algorithm can be considered to engage in discrimination, borrowing a definition by Lippert-Rasmussen [10].

Next, we will discuss foundational research on algorithmic bias that took place within the Data Mining (DM) community [9] and later among Machine Learning (ML) researchers. While ensuring fairness, accountability, and transparency have been considered key strategic elements in Information Retrieval [7], algorithmic bias concerns have been much less studied than in DM and ML. Studying biased ranking algorithms requires to shift the perspective from the people issuing a search, to the organizations and people that are represented by the items being searched. Specifically, we will seek for a *sufficient presence*, a *consistent treatment*, and a *proper representation* of items, particularly those belonging to protected or disadvantaged groups [5].

In the main part of the talk, we will overview some key results on fair rankings introducing methods to measure fairness in a ranking [13, 14] concepts such as fairness of exposure [12], and algorithms to ensure rankings are fair [2, 6], among others. We will also describe new methods for transparency and explainability in ranking.

Finally, we will describe ways of moving forward, avoiding the paralysis of having multiple competing fair ranking definitions [11], and show examples of detection and mitigation of biased rankings in practice.

ACKNOWLEDGMENTS

The author thanks La Caixa project LCF/PR/PR16/11110009 for partial support.

This talk is partially based on a keynote at the CIKM 2018 Workshop on Data and Bias [5] and a talk at the HUMAINT 2018 Workshop [4], both by the same author.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* 23 (May 2016).
[2] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. *arXiv:1805.01788 (pre-print)* (2018).
[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.
[4] Carlos Castillo. 2018. Algorithmic Discrimination at 1st HUMAINT Workshop. In *Assessing the impact of machine intelligence on human behaviour*. Joint Research Centre, 23–26.
[5] Carlos Castillo. 2018. Fairness and Transparency in Ranking. *ACM SIGIR Forum* 52, 2 (Dec 2018), 64–71. http://chato.cl/slides/DAB_2019_fairness_transparency.pdf.
[6] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with fairness constraints. In *Proc. of ICALP*. 28:1–28:15.
[7] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker (editors). 2018. Report from the Third Strategic Workshop on Information Retrieval(SWIRL). (2018).
[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015, 1 (2015), 92–112.
[9] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proc. of SIGKDD (Tutorials)*. ACM, 2125–2126. http://francescobonchi.com/algorithmic_bias_tutorial.html.
[10] Kasper Lippert-Rasmussen. 2014. *Born free and equal?: a philosophical inquiry into the nature of discrimination.* Oxford University Press.
[11] Arvind Narayanan. 2018. Tutorial: 21 fairness definitions and their politics. https://www.youtube.com/watch?v=jIXIuYdnyyk.
[12] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proc. of KDD*. 2219–2228.
[13] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proc. of SSDB*. ACM, 22.
[14] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Mohamed Hajian, Saraand Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. (2017), 1569–1578.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316451

A Link-based Approach to Detect Media Bias in News Websites

Victoria Patricia Aires
Institute of Computing – Federal University of Amazonas (UFAM)
Manaus, AM
victoria.aires@icomp.ufam.edu.br

Fabiola G. Nakamura
Institute of Computing – Federal University of Amazonas (UFAM)
Manaus, AM
fabiola@icomp.ufam.edu.br

Eduardo F. Nakamura
Institute of Computing – Federal University of Amazonas (UFAM)
Manaus, AM
nakamura@icomp.ufam.edu.br

ABSTRACT

News websites are currently one of the main sources of information. Like traditional media, these sources can have a bias in how they report news. This media bias can influence how people perceive events, political decisions, or discussions. In this paper, we describe a link-based approach to identify news websites with the same political orientation, i.e., characterize the bias of news websites, using network analysis techniques. After constructing a graph from a few seeds with previously known bias, we show that a community detection algorithm can identify groups formed by sources with the same political orientation.

CCS CONCEPTS

• **Information systems → Web applications**; *Data extraction and integration.*

KEYWORDS

media bias detection; news analysis; network analysis

ACM Reference Format:
Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. 2019. A Link-based Approach to Detect Media Bias in News Websites. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316460

1 INTRODUCTION

Nowadays, due to the popularization of the Web and social media, news websites have become one of the main sources of information [5]. However, like in traditional media, these portals have a bias in how they report news. By using aspects such as selective omission and choice of words, each source conveys a different impression of a fact [8]. This may impact the way the audience perceives events, political decisions and discussions regarding several topics. Studies have shown, for example, that media bias can influence the outcome of elections [4]. Detecting media bias is very hard for humans because of the inherent subjectivity involved [14]. Therefore, developing automatic methods for this purpose is an interesting research direction.

In this context, studies have been developed to automate the process of detecting media bias in news websites, most of them focusing on analyzing the textual content of the webpages [5, 11, 12]. However, news webpages form a link structure that can be explored, possibly revealing relationships between news websites. The link structure has been successfully studied in problems such as web spam and search engine rankings [7, 10], revealing that similar webpages form groups on the Web.

In this paper, we present a preliminary study to characterize the bias of news websites, i.e., identify portals with the same political orientation, using a link-based approach. The focus is on politics because the bias of news websites was previously classified by fact checking websites such as Media Bias Fact Check [3]. In our approach, we construct a graph with data crawled from a few seeds whose bias is known. Then, we apply network analysis tools such as community detection algorithm and centrality measures. We analyze the resulted graph, showing that websites with the same orientation form groups on the Web through links established between them.

The remainder of the paper is organized as follows. Section 2 includes works that focused on media bias detection on websites. In Section 3, we describe the steps of our approach. In Section 4, we report our experiments and results. Finally, Section 5 contains our conclusions and opportunities for future work.

2 RELATED WORK

In this section, we describe studies that address problems such as bias detection and political orientation detection in online news.

Efron [6] introduced a method for estimating political orientation of hypertext documents using cocitation information employing a probabilistic model and evaluating the likelihood of cocitation between the documents and a few seeds, i.e., documents of known orientation. Results showed that the model outperformed lexically based classifiers such as naïve Bayes and SVM.

Dallmann et al. [5] performed an analysis to identify political orientation of German online newspapers. They proposed measures that indicate a bias towards a political party, including mentions to a party and the sentiment associated with them. The results showed that the analyzed newspapers have bias towards specific parties, which is consistent with the public perception.

Morstatter et al. [11] focused on identifying framing bias, where specific aspects of a story are reinforced over others. They studied the ability of a machine learning classifier to detect frames and polarity in sentences of a news corpus. The results showed that simple linguistic features and the use of *n*-grams performed best in finding frames in text.

Niculae et al. [12] proposed an unsupervised framework based on quoting patterns for analyzing how the media selects what to cover in an article. They applied this framework to a dataset of political news and presidential speeches. They showed that the main dimension of bias align with the ideological spectrum and

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316460

source type, exposing differences in how different sources portray reality.

These works are mostly focused on the text of the articles, analyzing aspects such as vocabulary. In this paper, we intend to use a link-based approach to identify news websites with the same political orientation. Our main innovation is using network analysis tools such as community detection to characterize groups with the same political bias, something unexplored by previous works.

3 METHODOLOGY

In this section, we describe our approach to detect the bias of a news website. Below, we discuss each step of the proposed approach.

3.1 Selecting News Websites

The first step is selecting websites to be seeds in a web crawling process. We need to know in advance the bias of the seeds. Thus, we used Media Bias Fact Check (MBFC) [3]. MBFC is a fact checking site that rates websites based on ideological bias and credibility of factual reporting. Their methodology is subjective, but based on a numeric scoring system to assign labels. The possible labels for political bias are: Left, Left Center, Center, Right Center and Right. They are assigned on domain level, so every article originating from the same source will have the same label. For each class of political bias, we choose four websites to represent them as our seeds. These websites are listed in Table 1.

Table 1: Websites selected as our seeds and their respective bias, as determined by Media Bias Fact Check (MBFC) [3].

Bias	Website
Left Bias	https://www.cnn.com/
	https://www.huffingtonpost.com/
	http://nymag.com/
	https://theintercept.com/
Left Center Bias	https://www.bbc.com/
	https://www.latimes.com/
	https://www.nytimes.com/
	https://www.washingtonpost.com/
Center	https://www.reuters.com
	http://apnews.com
	https://www.politico.com/
	https://www.desmoinesregister.com/
Right Center Bias	https://www.forbes.com/
	https://nypost.com/
	https://www.thetimes.co.uk/
	https://www.wsj.com/
Right Bias	https://www.foxnews.com/
	https://observer.com/
	https://dailycaller.com/
	https://www.thesun.co.uk/

3.2 Crawling Pages and Creating the Graph

After selecting the seeds, the second step is to implement the web crawler to collect the data required to construct the graph. We implemented the web crawler using Python 3.7 and Scrapy, a web crawling framework [13]. The websites listed in Table 1 were given as input to the crawler. As visiting these URLs, it identifies all the hyperlinks in the page, and visits these links next. The process continues recursively until the crawler reaches 4 levels deep from the starting page in the Web graph, or the list of links to visit ends.

When visiting a webpage, we use the links to create the graph. We add in a CSV file an entry representing the source and the destiny of an edge. The source is the domain of the webpage we are visiting, and the destiny is the domain of the webpage it links to. However, we don't add an entry where the source and the destiny are the same website, filtering internal links. Another filter is removing links to social networks and advertisements of major brands.

At the end of the process, the CSV file is populated with several entries representing connections between websites. In this file, each domain is a node, and each entry is an edge. Repeated entries increase the edge weight, resulting in a directed and weighted graph.

3.3 Analyzing the Graph

The third step is to analyze the graph previously generated. Here, we used Gephi, a network analysis software [1]. The graph, represented in the CSV file, is given as input to Gephi. Before the analysis, we applied some filters. First, we merged entities from the same newspaper. For example, Page Six is the entertainment column of New York Post. In our analysis, we consider them as part of the same publication. Second, we filtered nodes with degree lesser than 10. After this filter, a new graph is generated without the removed nodes and their links.

Subsequently, we applied a community detection algorithm called Louvain method. The algorithm identifies groups in the graph that are more connected between them than with the rest of the graph [2] and it is based only on network topology, i.e., considering only the link structure of the network. We also applied HITS (Hyperlink-induced Topic Search), an algorithm that measures the centrality of the nodes in the network using scores called hubs and authorities [10], to identify the most influential nodes in each community. We focused on hub scores, which indicates the nodes that make links with higher relevance.

4 EXPERIMENTS & RESULTS

In this section, we report the results obtained after applying our methodology. After the web crawling process, the resulted graph had 3274 nodes and 4782 edges. Below, we describe two experiments. In the first, we applied the filters we described in Section 3. However, we noticed that the graph contains some websites that weren't labeled by MBFC. So, in the second experiment we analyze the graph composed only by websites whose labels are known. After describing both experiments, we discuss our results.

4.1 Experiment 1: Applying Basic Filters

In this experiment, we applied the filters we described in our methodology. This resulted in a new graph containing 98 nodes and 743 edges, illustrated in Figure 1. Five communities were detected, represented by colors in the figure. Below, we describe the communities, considering unlabeled websites. Table 2a shows the number of websites of each bias that occurred in each community.

Table 2: Distribution of websites and biases in our experiments.

(a) Experiment 1.

	Left	Left Center	Center	Right Center	Right	Unlabeled
Community 1	11	17	7	4	1	9
Community 2	2	9	5	3	4	9
Community 3	-	2	-	-	-	1
Community 4	-	1	-	2	1	7
Community 5	-	-	-	-	-	3

(b) Experiment 2.

	Left	Left Center	Center	Right Center	Right
Community 1	9	8	5	2	1
Community 2	4	15	4	2	2
Community 3	-	4	-	3	2
Community 4	-	1	3	-	-
Community 5	-	1	-	2	1

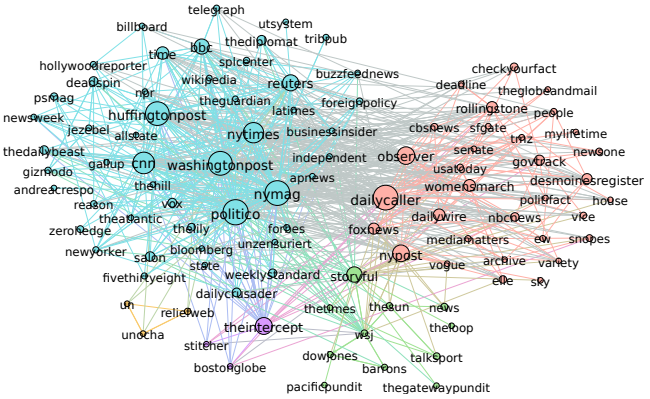


Figure 1: The resulted graph after applying our approach. The colors represent the communities detected and the size of the nodes corresponds to their hub scores.

Community 1 (blue). The main nodes, according to hub scores, are Politico, New York Magazine and Washington Post. These portals have, respectively, Center, Left and Left Center bias. The community is mainly composed by left-wing websites (Left and Left Center bias). Between the unlabeled websites, we highlight Daily Crusader, a junk news website, i.e., a portal that publishes low quality news using serious journalism to support their claims [9]. Storyful linked to this website in an article about questionable content. Daily Crusader is in this community because it links to credible sources such as New York Times, Wikipedia and Washington Post, which are also in this group.

Community 2 (red). In this community, the main nodes are The Daily Caller, New York Observer and New York Post. They have Right and Right Center bias, respectively. The community has almost the same number of left- and right-wing portals. However, as we mentioned earlier, right-wing nodes have most influence in this community. We believe the mixing occurred due to the tendency of these portals to criticize other of the opposite bias. Between the unlabeled websites, we highlight Women's March. The Daily Caller cited this website to criticize the movement and some members. On the other hand, Huffington Post had some articles communicating about the movement in a positive way. This shows how the sources report facts accordingly to their orientation. It also shows how right-wing portals criticize sources related to the left-wing.

Community 3 (purple). This was a small community, containing only The Intercept and The Boston Globe, two Left Center websites; and an unlabeled website that is an advertising of a small company. The presence of this website probably influenced the Louvain algorithm, which is greedy and based only on links, leading to an isolated community.

Community 4 (green). The main nodes in this community are Storyful, news.com.au and Wall Street Journal. The first is unlabeled; the second has Left Center bias, and the latter has Right Center bias. Most of the websites in this community are right-wing portals. Between the unlabeled, there are two websites classified by MBFC as Questionable Sources, i.e., sources that exhibits extreme bias, promotion of conspiracies and/or fake news [3]. These websites were linked by Storyful, which is in this community, in two articles about unreliable portals, acknowledging their suspicious nature.

Community 5 (yellow). Another small community, it is composed by unlabeled websites only. However, it's interesting to observe that they are humanitarian websites, such as United Nations and ReliefWeb. They have links only between them, showing their independence of news portals and therefore justifying the community.

4.2 Experiment 2: Filtering Unlabeled Websites

In this experiment, we wanted to check the impact, regarding to community detection, of removing unlabeled websites. The removal resulted in a new graph with 69 nodes and 575 edges. Five communities were detected, as in Experiment 1. Figure 2 illustrates the graph and the colors indicate the communities, and Table 2b shows the number of websites of each bias in each community. Similarly to Section 4.1, we describe the communities below.

Community 1 (blue). In this community, the main nodes are New York Magazine, Politico and Washington Post, respectively of Left, Center and Left Center biases. It contains mostly left-wing websites, with few right-wing sources, demonstrating a good grouping. We highlight that The Intercept and The Boston Globe, which had their own community in the previous experiment, were grouped with other left-wing websites after filtering unlabeled websites.

Community 2 (red). The main nodes in this community are The Daily Caller, Huffington Post and New York Times. The first has Right bias and the two latter, Left and Left Center bias, respectively. This community has majorly left-wing websites, and the hub scores also indicates that these are the most influential nodes. Again, the criticism of right-wing websites (linking to sources of the opposite bias) may have influenced in this grouping.

Community 3 (pink). Here, the main nodes are New York Observer, New York Post and Fox News, two of Right Center bias and one of

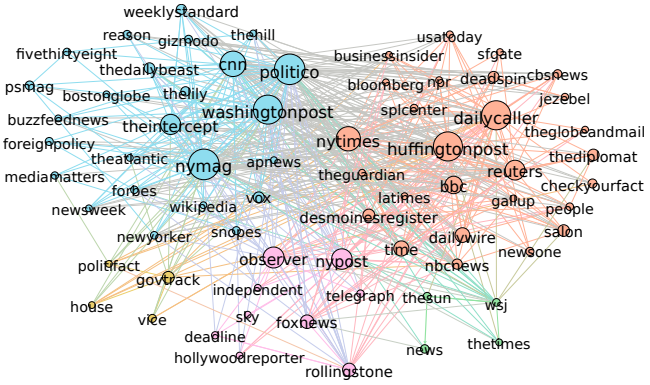


Figure 2: The graph after filtering unlabeled websites. The colors represent the communities detected and the size of the nodes corresponds to their hub scores.

Right. The community is mostly composed of right-wing websites. The websites of Left Center bias are mostly entertainment news portals, which can be interesting to the audience of the right-wing websites, justifying the links and, therefore, the community.

Community 4 (green). This community is very similar to Community 4 of Experiment 1, with the same nodes, excluding unlabeled websites. Like the other, this community has mainly right-wing websites and the main nodes are news.com.au, The Sun and Wall Street Journal.

Community 5 (yellow). This is a small community, however, it's interesting to observe that it's composed mainly of Center websites. The main nodes are GovTrack, House.gov and Politifact. The first two are related to the United States Congress, and the latter is a fact checking websites dedicated to U.S. politics. So, this community grouped the least biased sources that are related to American politics.

In both experiments, five communities were detected. In the first experiment we observed a mix in the communities, with most of the right-wing websites grouped with left-wing sources. In the second experiment, after removing unlabeled websites from the analysis, the detected communities were more discriminative. Each one of them had a majority belonging to one of the bias classes, namely: Left and Left Center (Community 1 and Community 2); Right and Right Center (Community 3 and Community 4); and Center (Community 5). We believe that these results indicate that a link-based approach can be a promising technique to characterize the bias of websites, leading to an interesting characterization that reflected what would be expected based on human judgement.

5 CONCLUSION & FUTURE WORK

In this paper, we presented a link-based approach to detect the bias of news websites. With only a few seeds of each bias class, we could construct a graph representing the connections between news portals. In this graph, we applied a community detection algorithm, based only on network topology, to identify the groups and check if they are composed of websites with the same political orientation.

The results showed that the communities are related to the bias of the websites, indicating that websites with similar biases tend to establish connections between them. We also noted that the communities reflected the common sense regarding the portals that were grouped together. Thus, a link-based approach can be effective to identify if a determined source is aligned with others.

However, since the results described are part of a preliminary study, there is room for improvements. In our experiments, some communities mixed sources of different bias due to links between them, established mainly by right-wing portals criticizing left-wing portals. Based on this, we intend to expand the study developing, in a future work, an approach to analyze the sentiment associated to the article or to the citation. This can be useful to determine if the websites have a similar point of view, therefore improving the link-based method.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and it is part of Process 88887.150732/2017-00 PGCI - Call 02/2015, held in cooperation by the Department of Computer Science and Engineering (CSE) at New York University (NYU) and by the Instituto de Computação (IComp) at the Universidade Federal do Amazonas (UFAM).

REFERENCES

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154
[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
[3] Media Bias Fact Check. 2019. The Most Comprehensive Media Bias Resource. Accessed January 24, 2019 from https://mediabiasfactcheck.com/.
[4] Chun-Fang Chiang and Brian Knight. 2011. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies* 78, 3 (2011), 795–820.
[5] Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media bias in german online newspapers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 133–137.
[6] Miles Efron. 2004. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. ACM, 390–398.
[7] Dennis Fetterly, Mark Manasse, and Marc Najork. 2004. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*. ACM, 1–6.
[8] Matthew Gentzkow and Jesse M Shapiro. 2006. Media bias and reputation. *Journal of Political Economy* 114, 2 (2006), 280–316.
[9] C Jensen. 2001. Junk Food News 1877-2000. *Phillips, P.(2001) Censored* 2001 (2001), 251–264.
[10] Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.
[11] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying Framing Bias in Online News. *ACM Transactions on Social Computing* 1, 2 (2018), 5.
[12] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 798–808.
[13] Scrapy. 2019. A Fast and Powerful Scraping and Web Crawling Framework. Accessed January 19, 2019 from https://scrapy.org/.
[14] Sevgi Yigit-Sert, Ismail Sengor Altingovde, and Özgür Ulusoy. 2016. Towards detecting media bias by utilizing user comments. In *Proceedings of the 8th ACM Conference on Web Science*. ACM, 374–375.

Workload Characterization of a Software-as-a-Service Web Application Implemented with a Microservices Architecture

Harold Aragon
Escuela Superior Politecnica del Litoral, ESPOL. Guayaquil, Ecuador.
haragon@espol.edu.ec

Samuel Braganza
Escuela Superior Politecnica del Litoral, ESPOL. Guayaquil, Ecuador.
sbraganz@espol.edu.ec

Edwin F. Boza
Escuela Superior Politecnica del Litoral, ESPOL. Guayaquil, Ecuador.
eboza@espol.edu.ec

Jonathan Parrales
Escuela Superior Politecnica del Litoral, ESPOL. Guayaquil, Ecuador.
jfparral@espol.edu.ec

Cristina L. Abad
Escuela Superior Politecnica del Litoral, ESPOL. Guayaquil, Ecuador.
cabadr@espol.edu.ec

ABSTRACT

We study the workload of an Online Invoicing application with clients in the Andean region in South America. The application is offered to clients with a Software-as-a-Service model, has a microservices architecture and runs on a containerized environment on a public cloud provider. The cloud application workload described in this paper can be used as part of a workload suite comprised of different application workloads, when evaluating microservices architectures. To the best of our knowledge, this is a novel workload in the web domain and it complements other workloads publicly available. Though we make no claim of the general applicability of this workload as a "microservices benchmark", its inclusion in evaluations could aid researchers and practitioners enrich their evaluations with tests based on a real microservices-based web application. Finally, we provide some insights regarding best-practices in microservice design, as a result of the observed workload.

CCS CONCEPTS

• **General and reference** → **Measurement**; • **Software and its engineering** → *Cloud computing*.

KEYWORDS

microservices; workload characterization; software-as-a-service

ACM Reference Format:
Harold Aragon, Samuel Braganza, Edwin F. Boza, Jonathan Parrales, and Cristina L. Abad. 2019. Workload Characterization of a Software-as-a-Service Web Application Implemented with a Microservices Architecture. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316466

1 INTRODUCTION

In a microservice architecture, an application is formed by a series of loosely coupled services, each performing a single functionality. These services are fine grained and communicate using lightweight protocols. Each microservice should be elastic and resilient, and have clearly defined APIs for ease of composability [7, 12, 17].

Microservices architectures have been increasingly adopted by enterprises [4, 20], mostly due to: (1) the success of the high-performance microservice-based applications of giants like Netflix, Amazon and eBay, and (2) the benefits of these architectures, which include enabling services that are independently developed and deployable, ease of horizontal scalability in Infrastructure-as-a-Service (IaaS) clouds, and support for efficient development team structures [19].

In the quest to increase the adoption of microservices architectures, many enhancements to the cloud infrastructure are being proposed. These modifications and enhancements seek to improve the performance, scalability and fault tolerance of microservice architectures under dynamic workloads (e.g., see [14, 16, 22]).

To better support these proposed enhancements, we need a better understanding of the types of workloads that microservice architectures impose on the computing infrastructure. Sadly, there is no freely available repository with microservice workloads.

We study and characterize the workload of a web application implemented with a microservices architecture. The workload comes from an invoicing-as-a-service application [8]. The results of our workload characterization can help others better understand how real microservice architectures work. Furthermore, our results can be used to parameterize simulations and benchmarks used to evaluate proposed infrastructure performance enhancements. To the best of our knowledge, this is a novel workload in the web domain and it complements other workloads publicly available. In addition, we provide insights supporting existing best-practices in microservice design, as a result of the observed workload.

2 DATASET

We studied a set of traces provided by Dátil (https://datil.co), a company with clients in the Andean region of South America. Specifically, they provided us with traces recording all requests received by the microservices supporting their Invoicing-as-a-Service product, during February, March and the first 18 days of April (2017). Dátil is a Software-as-a-Service provider that enables businesses to perform their invoicing online, by connecting to Dátil's web application. The businesses are charged a monthly subscription, depending on pre-defined service tiers (e.g., based on number of transactions per month). This web application is implemented using a microservices

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316466

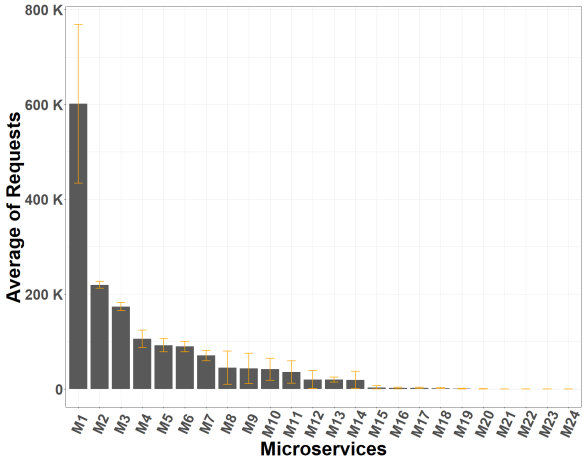


Figure 1: Average per-hour requests for each microservice, for the period between February and March 2017. The associated error bars correspond to one standard deviation.

architecture, and runs on a containerized environment (Docker) on a public cloud provider (Amazon Web Services).

The common log format of the traces included information about HTTP requests. For each request, the logs contain the request time, origin IP, destination IP and target microservice.

A preliminary analysis of this dataset was used to motivate prior work on simulation-based budgeting for web applications running on public cloud providers [8]. However, this analysis was limited to one microservice. In this paper, we extend this analysis by studying all the microservices supporting one web application.

An important limitation of our dataset is that the traces provided by Dátil do not contain information about dependencies between the microservices. For this reason, we do not study these dependencies or the workflows that they represent. Nevertheless, the information that can be analyzed without the dependency information is of use to experiments that seek to test aspects like load balancing, scalability, fault-tolerance, and performance.

Data processing. The data was preprocessed in Linux using standard command-line tools and converted to CSV (comma-separated values) files. The analysis and figure generation was done in R.

3 WORKLOAD CHARACTERIZATION

Unless otherwise noted, the results presented in this section correspond to the analysis of the complete dataset (Feb 01, 2017 through April 18, 2017). We focus our efforts to obtain statistical information that can be used by others to parameterize their benchmarks and simulators. We study common workload features like popularity, interarrival rates, and service times.

Microservice popularity. A common way to generate requests to target different "entities" (files, blocks, services, etc.) in a system, is to sample from the popularity distribution of these objects. This is referred to as the *Independence Reference Model (IRM)* [3], and it is frequently used for simulations [3] and benchmarking [13].

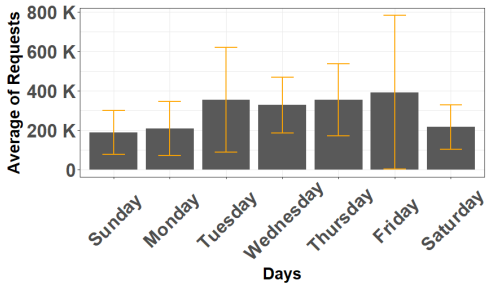


Figure 2: Average requests to M1 microservice per day of week. The error bars correspond to one standard deviation.

Figure 1 shows average per-hour requests for each microservice, for the period between February and March of 2017. The distribution is skewed, with a single microservice almost 3 times more popular than the second one, and a tail of unpopular microservices; however, the data is not well represented with a Zipf distribution (the gold standard when modeling popularity).

The data shows little variability in the hourly rate, except for the most popular microservice (M1). This behaviour should be considered for resource pre-allocation; if the service is elastically provisioned, tail latency due to cold starts triggered by scale-up decisions will increase [18]. The less popular microservice was only requested 23 times during the analyzed period.

Insight 1: Some microservices are extremely unpopular. Companies could consider migrating them to a serverless platform [1], e.g. AWS Lambda, instead of using a traditional container-based implementation; this approach could lead to significant savings without a noticeable decrease in performance [8].

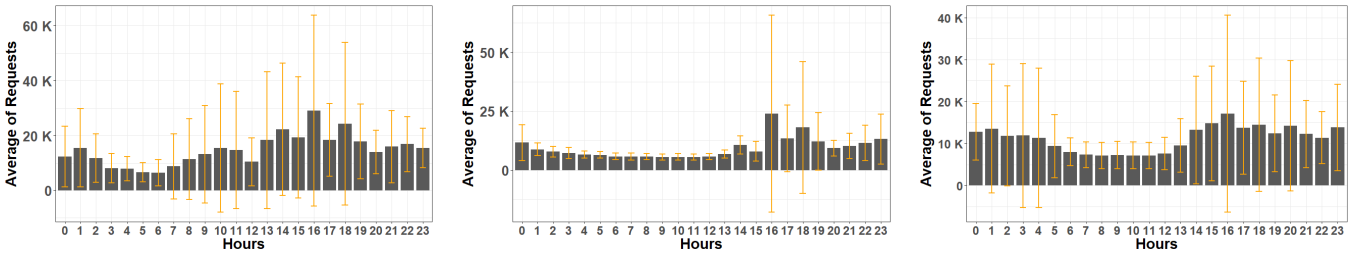
Request arrival rates. Figure 2 shows the average requests to the most popular microservice, received during each day of the week. The highest load is sustained Tuesday through Friday. This is the result of the higher commercial activity during these days.

Figures 3a– 3c show the average hourly load for the most popular service (M1), at day, week, and month levels. Load peaks at the 4-5pm slot, possibly due to some people leaving work at 4:30pm and increasing commercial transactions around this time. Least loaded hours are between 3am and 6am, when most businesses, other than pharmacies and gas stations, are closed. We can observe that the data from the third week of February is not representative of the hourly distribution observed at a monthly scale.

Insight 2: The high variability in request intensity at the daily and weekly levels, combined with the unpredictability of some of the variability, calls for the use of responsive auto-scalers that minimize costs while maintaining a target performance level.

Burstiness in arrivals. Figure 4 shows the burstiness of the requests received by M1 at different time scales: requests per second, per minute, per 10 minutes, and per hour. The load is very bursty, with peaks of high activity followed by plateaus with low load.

Insight 3: Requests are bursty at several levels, suggesting self-similar behavior; this has implications related to queuing theory at



(a) Per-hour requests in Tuesday days. **(b) Per-hour requests in 3rd week of February.** **(c) Per-hour requests in February.**

Figure 3: Average per-hour requests to the most popular microservice (M1). Error bars correspond to one standard deviation.

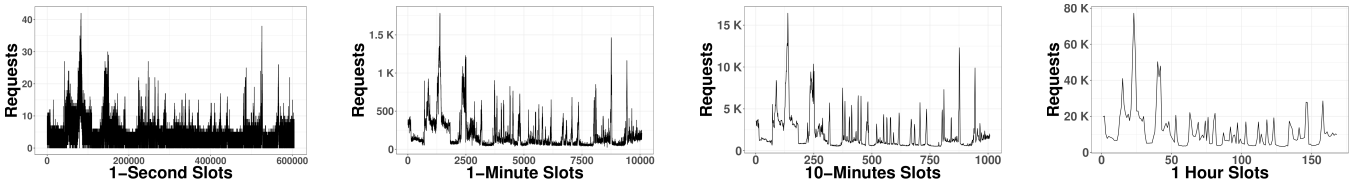


Figure 4: Requests in slots of different size in the first week of March for M1 microservice.

any place where requests may be queued (e.g., lost requests due to overflowing buffers).

Interarrival times. Figure 5 shows the cumulative distribution function (CDF) of the interarrival times for requests received by microservice M1. The tail, i.e. interarrivals greater than 10 seconds, was omitted to be able to show the most significant data in detail.

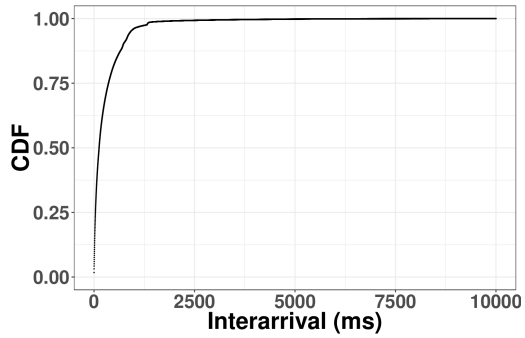


Figure 5: Interarrival distribution for M1 microservice.

Service Time. Figure 6 shows the backend processing time, which the load balancer defines as "The total time elapsed, in seconds, from the time the load balancer sent the request to a registered instance until the instance started to send the response headers." [5]. In other words, this records the time that it takes a microservice to process a request (i.e., the *service time*, using queuing theory terminology). We can observe that most microservices have a service time of less than 0.5 seconds. Furthermore, only two microservices, M8 and M9, have a service time that exceeds 3 seconds (on occasion); the largest observed service time was 60 seconds.

Insight 4: The combination of microservices with low and high response times in a single application calls for the use of diverse communication models (e.g., distributed message queues, pub-sub, and lightweight RPCs) that best suits each use case.

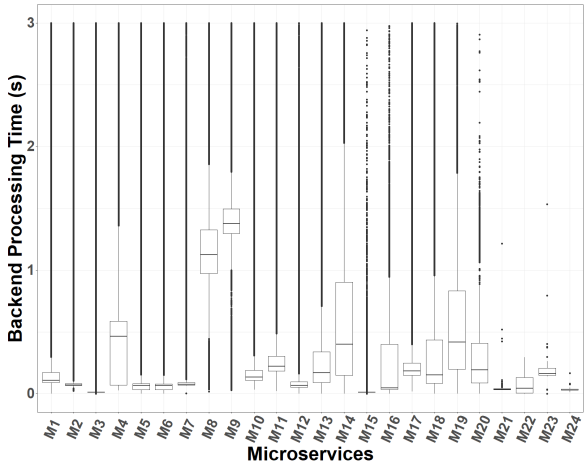
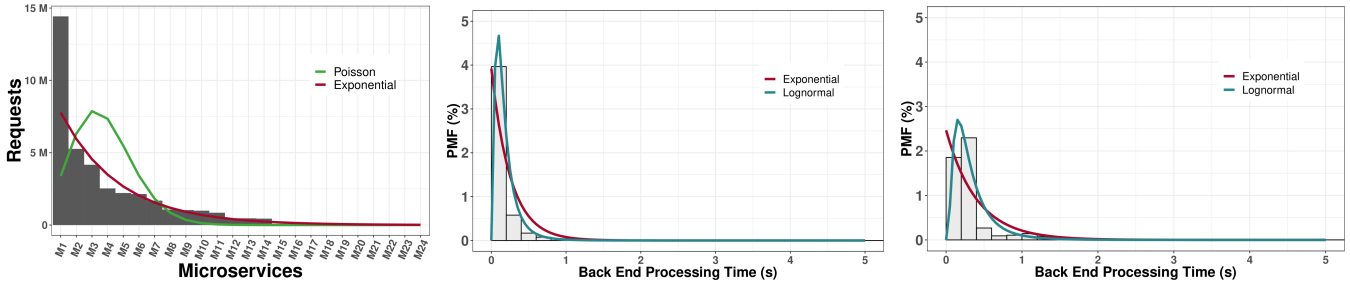


Figure 6: Boxplots of the microservice's service times.

Distribution Fitting. To facilitate the tasks of generating data for simulations and benchmarks, we provide the best fit of some of the studied data to common statistical distributions.

The graphical results are shown in Figure 7, while the Tables 1 and 2 quantify the difference between the observed data and the best fits, using the Kolmogorov-Smirnov distance. Their corresponding lognormal and exponential fitting parameters are $meanlog = -1.32$, $sdlog = 0.72$, and $rate = 2.46$ for M1; and $meanlog = -1.93$, $sdlog = 0.76$, and $rate = 3.92$ for M11. For microservice popularity, Poisson and exponential fitting parameters are $lambda = 0.73$ and $rate =$



(a) Best fit for microservices popularity (b) Best fit for service times of microservice M1 (c) Best fit for service times of microservice M11

Figure 7: Best fit for microservices popularity, and service times.

Table 1: Kolmogorov-Smirnov statistic for the best fit of the distribution of microservices popularity, to the exponential and poisson distributions.

	exponential	poisson
Kolmogorov-Smirnov statistic (D)	0.23493	0.21107

Table 2: Kolmogorov-Smirnov statistic for the best fit of the Service Times of microservices M1 and M11, to the exponential and lognormal distributions

	exponential	lognormal
M1	0.26217	0.20965
M11	0.27874	0.19094

0.26. We should note that the Zipf distribution was a very poor fit for our popularity data. The exponential distribution was not a good fit for the service times, but we nevertheless provide the best fit to this distribution, as it is frequently used in performance (queuing theory) models. Finally, due to space constraints we do not show the best fit of the interarrival times to the exponential distribution: $\lambda = 0.003573037$, with $D = 0.15421$ (Kolmogorov-Smirnov goodness of fit). For all presented data, we tried fitting to other common distributions, but do not present the results due to obtaining very poor results.

4 RELATED WORK

Workload characterization studies are important as they can be used as the basis for performance modeling, simulations and benchmarking [15]. In the past, others have studied in detail workloads from several related domains, like web servers [6], media servers [11], storage [2], MapReduce clusters [10], among others. In the microservices domain, a few application-based synthetic benchmarks exist, most notably, Acme Air [21], *μSET* [9] and Sock Shop[1]. Our work can be used to parameterize benchmarks like these, and also as the basis of simulation-based studies. However, a good evaluation should not be based on one single workload, so the need for public information and traces regarding real microservices workloads remains an open problem that hinders the evaluation of microservice architectures as well as emerging improvements to the infrastructure that supports them.

[1]https://microservices-demo.github.io/

5 CONCLUSIONS

We presented the results of a workload characterization study of a small Online Invoicing application implemented using a microservices architecture. The cloud application workload described in this paper can be used as part of a workload suite comprised of different application workloads to evaluate microservice architectures, or as the basis of simulation-based performance studies. We urge others to publish similar studies of real applications with microservices architectures, so that the community can evaluate their systems using a wide variety of representative workloads.

REFERENCES

[1] Cristina Abad, Edwin Boza, and Erwin van Eyk. 2018. Package-Aware Scheduling of FaaS Functions. In *Companion of ACM/SPEC ICPE*.

[2] Cristina Abad, Nathan Roberts, Yi Lu, and Roy Campbell. 2012. A storage-centric analysis of MapReduce workloads: File popularity, temporal locality and arrival patterns. In *IEEE IISWC*.

[3] Cristina Abad, Mindi Yuan, Chris Cai, Yi Lu, Nathan Roberts, and Roy Campbell. 2013. Generating request streams on Big Data using clustered renewal processes. *Performance Evaluation* 70, 10 (2013).

[4] Carlos Aderaldo, Nabor Mendonça, C Pahl, and P Jamshidi. 2017. Benchmark requirements for microservices architecture research. In *IEEE/ACM ECASE*.

[5] Inc. Amazon Web Services. 2018. Access Logs for Your Classic Load Balancer. In *AWS Documentation*. Amazon Web Services, 96.

[6] Martin F Arlitt and Carey L Williamson. 1996. Web server workload characterization: The search for invariants. *SIGMETRICS Perf. Eval. Rev.* 24, 1 (1996).

[7] SAMAN Barakat. 2017. Monitoring and Analysis of Microservices Performance. *Journal of Computer Science & Control Systems* 10, 1 (2017), 19.

[8] Edwin Boza, Cristina Abad, Mónica Villavicencio, Stephany Quimba, and Juan Plaza. 2017. Reserved, on demand or serverless: Model-based simulations for cloud budget planning. In *IEEE Ecuador Technical Chapters Meeting (ETCM)*.

[9] Antonio Brogi, Andrea Canciani, Davide Neri, Luca Rinaldi, and Jacopo Soldani. 2017. Towards a reference dataset of microservice-based applications. In *International Conference on Software Engineering and Formal Methods*. Springer.

[10] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz. 2011. The Case for Evaluating MapReduce Performance Using Workload Suites. In *IEEE MASCOTS*.

[11] L. Cherkasova and Minaxi Gupta. 2004. Analysis of enterprise media server workloads: Access patterns, locality, content evolution, and rates of change. *IEEE/ACM Transactions on Networking* 12, 5 (2004).

[12] Ghyzlane Cherradi, Adil El Bouziri, Azedine Boulmakoul, and Karine Zeitouni. 2017. Real-time hazmat environmental information system: A micro-service based architecture. *Procedia Computer Science* 109 (2017).

[13] Brian F Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. 2010. Benchmarking cloud serving systems with YCSB. In *ACM SoCC*.

[14] Maria Fazio, A Celesti, R Ranjan, C Liu, L Chen, and M Villari. 2016. Open issues in scheduling microservices in the cloud. *IEEE Cloud Comp.* 3, 5 (2016).

[15] Dror G Feitelson. 2002. Workload modeling for performance evaluation. In *IFIP Intl. Symp. Computer Performance Modeling, Measurement and Evaluation*.
[16] Ana Klimovic, Yawen Wang, Patrick Stuedi, Animesh Trivedi, Jonas Pfefferle, and Christos Kozyrakis. 2018. Pocket: Elastic ephemeral storage for serverless analytics. In *USENIX Symp. Op. Sys. Design and Impl. (OSDI)*.
[17] Holger Knoche. 2016. Sustaining runtime performance while incrementally modernizing transactional monolithic software towards microservices. In *ACM/SPEC ICPE*.
[18] Wes Lloyd, Shruti Ramesh, Swetha Chinthalapati, Lan Ly, and Shrideep Pallickara. 2018. Serverless computing: An investigation of factors influencing microservice performance. In *IEEE IC2E*.
[19] Genc Mazlami, Jurgen Cito, and Philipp Leitner. 2017. Extraction of microservices from monolithic software architectures. In *IEEE Intl. Conf. Web Services (ICWS)*.
[20] Takanori Ueda, Takuya Nakaike, and Moriyoshi Ohara. 2016. Workload characterization for microservices. In *IEEE IISWC*.
[21] T. Ueda, T. Nakaike, and M. Ohara. 2016. Workload characterization for microservices. In *IEEE IISWC*.
[22] Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. 2015. Large-scale cluster management at Google with Borg. In *European Conf. Comp. Sys. (EuroSys)*.

A Constellation of Horrors: Analysis and Visualization of the #Cuéntalo Movement

María Soledad Bucalo
Luz Calvo
Fernando Cucchietti
David García Povedano
Artur García-Sáez
Barcelona Supercomputing Center (BSC)
Barcelona, Spain

Juan Felipe Gómez
Camilo Arcadio González
Guillermo Marín
Patricio Reyes
Feliu Serra
Diana Fernanda Vélez
Barcelona Supercomputing Center (BSC)
Barcelona, Spain

Irene Meta
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Institute for Advanced Architecture of Catalonia (IAAC)
Barcelona, Spain

ABSTRACT

In this work, we analyze content and structure of the Twitter trending topic #cuentalo with the purpose of providing a visualization of the movement. A supervised learning methodology is used to train the classifying algorithms with hand-labeled observations. The methodology allows us to classify each tweet according to its role in the movement.

KEYWORDS

visualization; machine learning; supervised learning; social networks

ACM Reference Format:
María Soledad Bucalo, Luz Calvo, Fernando Cucchietti, David García Povedano, Artur García-Sáez, Juan Felipe Gómez, Camilo Arcadio González, Guillermo Marín, Patricio Reyes, Feliu Serra, Diana Fernanda Vélez, and Irene Meta. 2019. A Constellation of Horrors: Analysis and Visualization of the #Cuéntalo Movement. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316459

1 INTRODUCTION

The #cuentalo (hashtag for *cuéntalo*, which means *tell it* in Spanish) movement started on April 2018, inviting women to share on Twitter their personal experiences of sexual aggression. The movement was triggered in Spain with the news about the court decision in the so called *wolf pack* case [1]. In a few days, it generated more than 2.5 million tweets and retweets with narrations told by their protagonists.

The tweets go from the uncomfortable to the unbearable, stories in first person occasionally mixed with a woman speaking in the name of another one because she doesn't have a computer, or because she doesn't dare to tell her story, or because she was murdered.

Most cases of sexual aggression go unreported [6]. Many of the tweets and users involved in this viral phenomenon mentioned that this was the first time in their lives they talked about it. A related phenomenon was the #MeToo movement [11], although with the subtle difference that in Spanish, #MeToo induces a sense of appealing to be similar to the celebrities that started the movement, while #cuentalo is a more anonymous call asking women to end the silence and tell what happened to them, implicitly assuming that most have experienced something similar. In order to bring attention to the topic, to support the debate around it, and to help create and maintain a new collective memory, we conducted an analysis of the data produced by the movement.

In this research work we propose a methodology to analyze the content and structure of the Twitter trending topic #cuentalo which serves as the basis for a visualization tool that uncovers information related to the problem of sexual aggression. The analysis results in new perspectives and information about the frequency, types, and other details of sexual aggression.

2 RELATED WORK

Due to the similar nature of the movement, special attention deserves the literature about #MeToo. In [11], authors study common words, semantic relationships, sentiment analysis from content found on both Twitter and Reddit. Their results focus on the differences in content regarding these two platforms.

Concerning data visualization, most of the references found are data-visualization projects which analyze #MeToo and display the results on a website [5, 7]. Prior to the #MeToo movement, Jain [10] presents descriptive statistics about rape in India, with data from the justice system.

Regarding #cuentalo, with a similar dataset, in *Contando cómo se difundió el #cuéntalo* [4] the author performs a social network analysis to identify the topology of the network defined by the relationships between users (follower) or tweets (re-tweets).

3 METHODOLOGY

3.1 Dataset description

The raw dataset corresponds to the tweets published between April 27 and May 12, 2018 (except 4 days) with the hashtag #cuentalo.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316459

This dataset contained 2.1 million tweets. It was collected by members of the *Asociación de Archiveros-Gestores de Documentos de Cataluña (AAC-DG)*. The collection was missing the days: May 29, June 4, June 6, and June 11. For these days, we were able to recover the tweets with original content (no retweets), bringing the final dataset to 2.75 million tweets in total. Although less than 3% of tweets had proper geolocation information, by reading and parsing the *location* field we were able to locate city or country information for roughly half of all the tweets in the dataset. There were over 160 thousand tweets with content written by users, while the rest (2.6 million) were retweets. Although retweets were crucial for the virality of the movement, here we focus only on those with content, that we denote as *original*.

3.2 Classification techniques

3.2.1 Training database. The dataset [12] used for training purposes deserves a special mention. It consists of around 10632 randomly chosen original tweets categorized by volunteers specifically for this project. This corresponds to 6.6% of the original tweets. Only tweets in Spanish were categorized, which helped standardize the training dataset, and was the language of the vast majority of the tweets.

The categorization process was as follows: Each volunteer received a table, one row per tweet and one column per category or label, where each volunteer marked *yes/no* if the tweet content corresponds to the labels in **what** (see description below), except for the category **who**, where the volunteer chose between 5 possibilities. For each tweet, the volunteer had access to (1) the tweet id, (2) the tweet user name and (3) the content of the tweet.

The categories and labels were the following:

- **who**: (mutually exclusive choices: *1–5*) the writer of the tweet either:
  1. tells something about himself/herself.
  2. tells something about someone else.
  3. tells something supporting the movement.
  4. tells something not related to the movement.
  5. tells something against the movement.
- **what**:
  - **murder**: (*yes/no*) the tweet describes a murder
  - **rape**: (*yes/no*) the tweet states a rape or attempted rape
  - **sexual assault**: (*yes/no*) the tweet describes a sexual assault (but not situations that belongs to the category **rape**)
  - **abuse**: (*yes/no*) the tweet states abuse.
  - **harassment**: (*yes/no*) the tweet discusses about a non-physical harassment situation
  - **fear**: (*yes/no*) the tweet explicitly describes fear
  - **disgust/sadness/anger**: (*yes/no*) the tweet explicitly describes disgust/sadness/anger, only if **who** is different than 4.

Among the 10.632 categorized tweets, 31% correspond to testimonies written in first person (i.e. the answer to **who** is 1), 8.9% in third person or on behalf of someone else (**who** is 2), 40.2% are supportive tweets (**who** is 3), 3.1% correspond to tweets against the movement (**who** is 5), and 16.7% are tweets not related to the movement (unclassifiable) (**who** is 4). Percentages are depicted in Figure 1 and Figure 2.

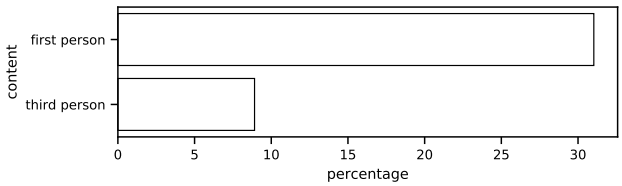


Figure 1: Type of content in original tweets according to the writer

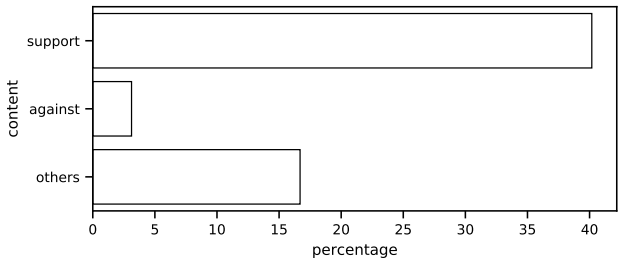


Figure 2: Type of content in the original tweets according to the support to the movement

Among the tweets with testimonies (written in either 1st or 3rd person), 3.9% talk about a murder, 5.6% about rape, 11.2% about a sexual assault, 6.3% about abuse, 14.2% harassment, 11.8% mention fear, and 19% disgust/sadness/anger (See Figure 3).

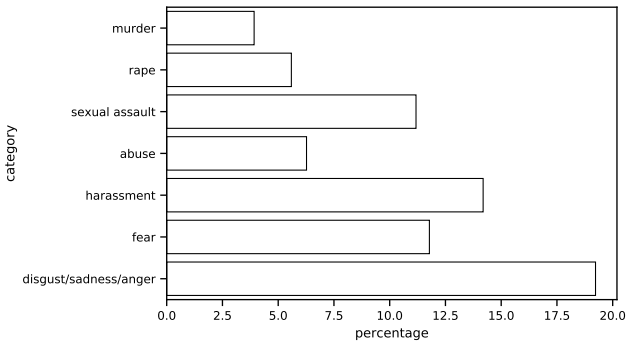


Figure 3: Type of content in original tweets according to the situation and sentiment

3.2.2 Classification process. The cost of classifying tweets manually is very large and can be imprecise because of human biases. Instead of cross-checking results between classifiers, we set some strict rules to help homogenize the classification criteria, which in our case turned out to be the most cost-effective approach. For example, the tweets containing images, text screen-captures, or written in a language other than Spanish were categorized as not related to the movement (**who**=4).

We did not cross-check the classification from different volunteers. However, all of them reached very similar percentages of the

different categories, hence we assume that the classification was precise enough.

Notice that our classification criteria might be seen as not sufficiently specific to describe correctly or in sufficient detail such a dramatic topic, but follows a critical discussion trying to balance important factors like frequency or duration of the aggression against technical limitations.

3.2.3 Supervised learning. Due to the properties of the dataset, we decided to use a supervised learning methodology in order to train the algorithms. We settled on a deep neural network multi-output classifier built using the Keras API to Tensorflow [3]. The version of the Keras package used is the 2.1.5.

3.2.4 Aggregation of outputs and cross validation. The complexity of the natural language used in tweets, such as spelling mistakes or occasional mixed languages, makes the multi-output classification even hard for humans. In order to improve the accuracy we had to narrow down the multi-output problem, and ended up with a simplification of the categories as follows: The answers for **who** were re-grouped as **testimony**/**support**/**others**: a tweet belongs to **testimony** if **who** $\in \{1, 2\}$, to **support** if **who**=3 or to **others** if **who** $\in \{4, 5\}$. In the same way, the answers for **what** were re-grouped as **physical**/**non-physical**/**others**: a tweet belongs to **physical** if any of the labels **murder**, **rape**, **sexual assault**, **abuse** was answered as *yes*; it belongs to **non-physical** if any of **harassment**, **fear**, **disgust/sadness/anger** was answered as *yes*; finally, it belongs to **others** if none of the labels in **what** were affirmative. By aggregating groups in this way we made the learning task easier for the learning algorithm, improving from an accuracy of 56% on the 5 output **who** classification to 75% on the 3 output classification problem. The re-grouping of **what** improved the accuracy from 56% on the 7 multi-output classification to a 68% accuracy on the grouped 3 categories (See section 3.2.5).

For the cross-validation we used a 60%/20%/20% split for the training, validation and testing sets respectively.

3.2.5 Neural network architecture. We tested different architectures, and here we report two of them. The first architecture A1, has the pretrained GloVe [14] word embedding, then a *dense* layer, an *activation* layer, and finally a *flatten* layer and a *dense* layer. The second architecture A2, was used to perform the final classification. It starts with the GloVe word embedding. After a dropout layer to regularize the embedding signal, it uses a convolutional layer with its respective maxpooling layer to reduce the dimensionality of the data. Next, long-short term memory layer, with a dropout layer, dense layer and a batch normalization layer in order to avoid over fitting. Finally, the dimensionality was reduced using another dense layer with a final softmax output layer for the multioutput classification task.

Table 1 shows a classification summary of the test set made using sklearn package's classification report method [8, 13]. It shows the precision for each category predicted, the recalls or true positive rate and the support or number of observations that have been hand classified as belonging to those classes. Table 2 shows the accuracy per each architecture (A1, A2) and category (who,what) considered.

Table 1: Out of sample classification summary

		precision		recall		support
		A1	A2	A1	A2	
who	testimony	0.80	0.70	0.88	0.92	772
	support	0.79	0.84	0.76	0.65	836
	others	0.64	0.73	0.56	0.65	392
	avg/total	0.73	0.76	0.71	0.75	2,000
what	physical	0.62	0.77	0.61	0.63	437
	non-physical	0.53	0.55	0.44	0.79	710
	others	0.65	0.82	0.75	0.60	853
	avg/total	0.60	0.71	0.61	0.67	2,000

Table 2: Accuracy according to the models and categories

	A1	A2
who	0.77	0.75
what	0.63	0.68

As we can see in Tables 1 and 2, there is no significant improvement between A1 and A2. This is partly due to the GloVe word embedding being an important part of the process.

We tried other variations of the architecture, such as using bidirectional LSTMs or stacking multiples LSTMs with their respective dropout layers, but the learning algorithm tended to overfit when there were too many parameters to tune considering the small amount of observations of the training dataset.

4 RESULTS

The results (See Tables 1 and 2) of the methodology above allows us to classify each tweet in two ways. First, into three mutually exclusive categories: Testimony, support and others. Second, the three classes regarding what the tweet was about: Physical content, non-physical content or others (See section 3.2.4). Each tweet was given the probability of belonging to each one of these categories. We show in Figure 4 the resulting classification using a Ternary plot [9]. The plot shows each classified tweet as a dot where the probability assigned to each of the three categories is represented as one minus the distance to one of the vertices of the triangle. Then, an almost perfect classification would show all dots near the vertices. Figure 4 shows relatively few dots near the center, which would indicate confusion.

5 VISUALIZATION

After the tweets were classified, a visualization that uncovers information related to the problem of sexual aggression was designed with the aim of informing the general audience about the movement, making the content easily available. After several discussions and prototypes [2], we decided that the resulting visualization must satisfy the following statements:

- All the tweets must be represented individually.
- Every tweet must have the same weight.
- The visualization must be flexible enough to receive new tweets.

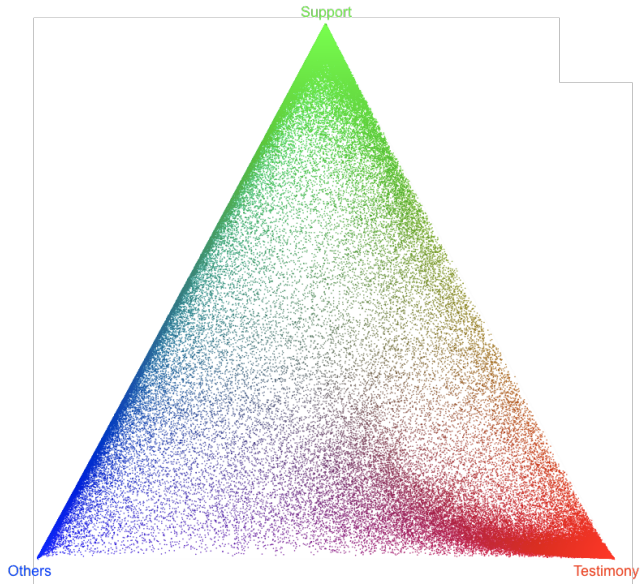


Figure 4: Ternary plot for the classification of the original tweets into three main categories.

Taking into account the statements above, and the classification depicted in Figure 4, we decided to classify the tweets with respect to their role in the campaign.

In the resulting visualization (See Figure 5) all the tweets are depicted in a radial distribution, where the angle corresponds to the time when the tweet was published. The radius, which is the distance with respect to the center, is related to the role of the tweet in the movement. The inner, red ring corresponds to testimonial tweets, surrounded by the outer pink ring, which corresponds to supporters. Out of the outer pink ring there are the tweets against or not related to the movement. The online visualization can be seen in the website of the project https://www.bsc.es/viz/cuentalo/.

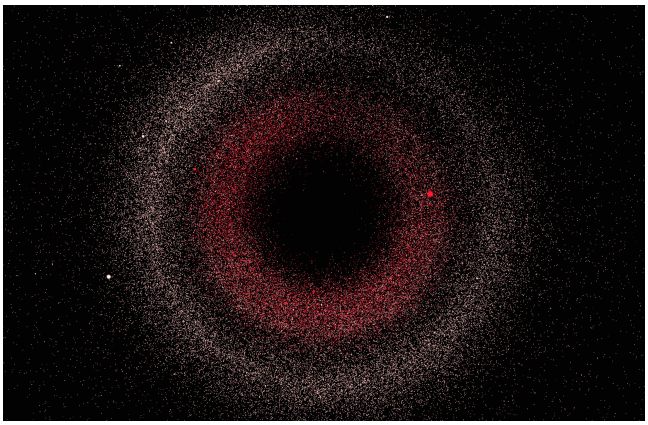


Figure 5: Snapshot of the final visualization.

6 CONCLUSIONS AND FUTURE WORK

We have proposed a methodology which classifies tweets with respect to their role in the #cuentalo movement. A visualization was delivered as a mean to present the general audience with information about sexual aggression in a more appealing manner.

As a byproduct, we have provided a dataset [12] of 10 thousand #cuentalo tweets classified by volunteers which can be useful for future studies of the movement.

As a future work, we are interested in an in-depth data analysis to determine the profile of the protagonists of the tweets. Moreover, we would like to compare the tweets between #cuentalo and the #MeToo movement.

7 ACKNOWLEDGMENTS

The authors would like to thank María Coto and Laura Gutiérrez for helpful discussions.

REFERENCES

[1] James Badcock. 2018. Fury over Rape Case Sparks Spanish #MeToo. (May 2018). https://www.bbc.com/news/world-europe-43960647
[2] bscvizteam. 2018. #Cuéntalo – La Visualización – The BSC Viz Corner. http://www.bsc.es/viz/corner/?p=210&lang=es#comments
[3] François Chollet et al. 2015. Keras. (2015).
[4] Mariluz Congosto. 2018. Contando Cómo Se Difundió El #cuéntalo. http://www.barriblog.com/2018/05/contando-se-difundio-cuentalo/?lang=en
[5] Valentina D'Efilippo. 2018. The Anatomy of a Hashtag — a Visual Analysis of the MeToo Movement. https://medium.muz.li/the-anatomy-of-a-hashtag-a-visual-analysis-of-the-metoo-movement-ba4ecf9b130b.
[6] Bonnie S. Fisher, Leah E. Daigle, Francis T. Cullen, and Michael G. Turner. 2003. Reporting Sexual Victimization To The Police And Others: Results From a National-Level Study of College Women. *Criminal Justice and Behavior* 30, 1 (Feb. 2003), 6–38. https://doi.org/10.1177/0093854802239161
[7] Lazaro Gamio and Haley Britzky. 2018. What to Know about All of the Men Facing Sexual Misconduct Allegations. https://www.axios.com/sexual-misconduct-allegations-since-weinstein-6da885ce-83a5-40bc-9fae-2b1b65f9c68e.html.
[8] Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative Methods for Multi-labeled Classification. In *Advances in Knowledge Discovery and Data Mining*, Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 22–30.
[9] Marc Harper et al. 2015. Python-Ternary: Ternary Plots in Python. (Dec. 2015). https://doi.org/10.5281/zenodo.34938
[10] Aditya Jain. 2018. Rape in India. https://adityajain15.github.io/Rape_In_India/.
[11] Lydia Manikonda, Ghazaleh Beigi, Subbarao Kambhampati, and Huan Liu. 2018. #metoo Through the Lens of Social Media. In *Social, Cultural, and Behavioral Modeling (Lecture Notes in Computer Science)*, Robert Thomson, Christopher Dancy, Ayaz Hyder, and Halil Bisgin (Eds.). Springer International Publishing, 104–110.
[12] pareyesv. 2019. BSCCNS/cuentalo-dataset. https://doi.org/10.5281/zenodo.2585526
[13] F. Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

Bitcoin Price Prediction Through Opinion Mining

Germán Cheuque Cerda
gacheuque@uc.cl
Pontificia Universidad Católica de Chile
Santiago, Chile

Juan Reutter de La Maza
jreutter@ing.puc.cl
Pontificia Universidad Católica de Chile
Santiago, Chile

ABSTRACT

The Bitcoin protocol and its underlying cryptocurrency have started to shape the way we view digital currency, and opened up a large list of new and interesting challenges. Amongst them, we focus on the question of how is the price of digital currencies affected, which is a natural question especially when considering the price roller-coaster we witnessed for bitcoin in 2017-2018. We work under the hypothesis that price is affected by the web footprint of influential people, we refer to them as crypto-influencers.

In this paper we provide neural models for predicting bitcoin price. We compare what happens when the model is fed only with recent price history versus what happens when fed, in addition, with a measure of the positivity or negativity of the sayings of these influencers, measured through a sentiment analysis of their twitter posts. We show preliminary evidence that twitter data should indeed help to predict the price of bitcoin, even though the measures we use in this paper have a lot of room for refinement. In particular, we also discuss the challenges of measuring the correct sensation of these posts, and discuss the work that should help improving our discoveries even further.

KEYWORDS

Bitcoin, Recurrent Neural Networks, Twitter, Sentiment Analysis, Price Prediction.

ACM Reference Format:
Germán Cheuque Cerda and Juan Reutter de La Maza. 2019. Bitcoin Price Prediction Through Opinion Mining. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316454

1 INTRODUCTION

The release of the Bitcoin Protocol [22] unleashed bitcoin as the world's first decentralized cryptocurrency. While it took some time to gain traction, in the last few years a huge number of alternate cryptocurrencies have appeared, and they have become popular to the point that almost every financial agent has at least considered investing in them. At the time of writing, the price of bitcoin has stabilized around 3500 USD, after an all time high of almost 20000 USD just two years ago. But while this price bubble is now understood to have burst –at least partially–, some other concerns about the condition of the bitcoin market are still not well understood.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316454

One of the most recurrent question about the bitcoin market has to do with its volatility: what factors influence the variation in bitcoin price? There has been some research that links bitcoin price with google searches [24], which is remarkable because other comparable assets such as gold are known to have very little correlation with this indicator. However, in conversation with people in the cryptocurrency market we were suggested another indicator: tweets and posts from a number of international influencers.

The goal of this research is to validate the hypothesis that the price of bitcoin is affected by the web footprint of popular actors in the international cryptomarket. Preliminary research has told us that the amount of digital material does not appear to be correlated with bitcoin price, so we cannot settle for a simple answer like in [24], But we can do more: what if positive messages affect the price in a positive way, and negative messages in a negative way? This idea gives us a clear map of what needs to be done: First, gather tweets by the most influential users in the cryptocurrency world. Next, analyze which of these messages speak positively of bitcoin and which ones do not, and then show that this data actually does affect bitcoin's price.

Influential users can be selected in a greedy way, starting from some well known popular users, and we can use sentiment analysis to classify the tweets of these users. But how can we show that this data affects the price? We proceed as in [1] and [2], and compare the prediction capabilities of a model that uses only the information of bitcoin's price in the past with a model that incorporates this external information into the time series.

In this paper we show promising preliminary results, showing that incorporating twitter data can lead to better forecasting models. However, we believe there is still much room for improvement, as there are many different ways to follow our general map, and we have only studied a small number of them. The study is presented as follows: in Section 2 we deepen our objectives, present the data we are working with and explain the type of prediction we will focus in the rest of the paper. Section 3 the architecture of our model, discusses the needed of an appropriate scaling and presents the results of the realistic problem of predicting price one day in future, with their new complications. Section 4 introduces the opinion information obtained from twitter and how their inclusion impacts on our results. Finally, Section 5 presents our conclusions and the proposed further work.

2 ON PREDICTING PRICE OF MONETARY GOODS

Predicting the price of monetary goods is a problem that has been thoroughly studied for several decades and from a series of different areas. As such, it is almost impossible to give a full overview of all the different versions of these problems, or the techniques used to

solve them.

However, we do have one important aspect to discuss: the problem of predicting the immediate price of monetary goods that are subject to a high volume of transactions can be dealt with (up to a very reasonable error) standard machine learning techniques. For the case of Bitcoin, in Section 2.2 we provide a neural network that can effectively predict the price of bitcoin in the next 5 minutes, and we test it to show that the error is indeed quite small [1].

For this reason we focus instead on other problems which cannot be directly dealt with standard techniques. One problem is predicting the price of Bitcoin after 24 hours of reading the last input, and the other problem has to do with training a model that can properly simulate bitcoin prices for a long period of time. We define these in more detail in Section 2.3

2.1 Samples and range of prediction

We draw Bitcoin price data from bitcoincharts.com, and focus on the interval between July 1st and December 30th over 2018. The data contains information about the opening and closing prices of Bitcoin on intervals of 5 minutes, giving us a total of 58451 samples. In all our models we use the first 37339 data points as training data[2], and use the remaining to test our predictions. The Bitcoin price evolution over the referred dates is presented in Figure 1.

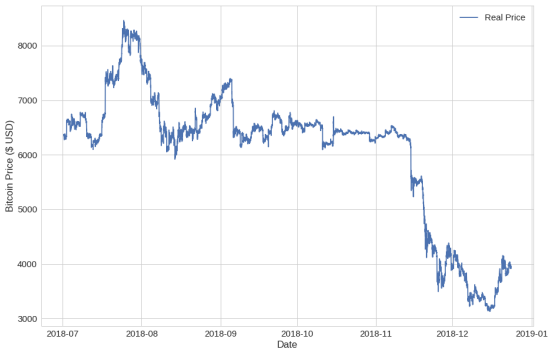


Figure 1: Bitcoin price time series.

As everyone in industry knows, the price of Bitcoin is still far from reaching a stable behavior. And while the high frequency of registries is definitely good news for learning models predicting its price, it also carries a curse: the large number of variations can be difficult to learn, and occasional points with high variablity would tend to be magnified by the necessary scaling we must perform on the data. This means that predictors may either choose to focus on these high variability changes, and therefore loosing constant small changes, or focus on small changes and treating big raises or drops as outliers. A way to solve this noisy effect without losing our high frequency chart is to introduce a new quantity, we refer to the *moving average* measure, defined by equation 1

$$MA = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{1}$$

where p_i is the current price at date i and N indicates the number of samples that are considered on average. The idea behind this measure is to work with a constant window of size N and consecutively move their bounds to get a continue measure of how the price varies on average. Figure 2 show us how moving average captures the variability of the price with a smoother approximation. We remark that moving average is one of the most used metrics in stock exchange markets, to the point that its inclusion is considered folklore.

Figure 2: Comparison between price and moving average 1-day measure.

2.2 Predicting the next data point.

If all we want is to predict the next data point in the series, we can obtain reasonable prediction by using state of the art neural networks. Many studies have addressed the problem of forecasting using LSTM in the most diverse fields. Some studies have worked on e-commerce [5], other have made the same analyzing the weather [6], even studying the human blood pressure [7]. Other studies have made important advances in popular computer tasks, such as question Answering [8], [9] or speech recognition [10] and even in the development of autonomous vehicles [11] with the use of LSTM.

We chose to build a Recurrent Neural Network based in LSTM. Our network consist of five layers, each one of 30 neurons, except for the first one, with 288 neurons. Each one with their respective dropout regularization, the final two layers are fully connected to retrieve a single value for prediction. Further details about configuration and design decision can be found in Section 3. The network is fed with a moving window N of 288 data points, i.e., one full day of observations, on the other hand, the output is trained to predict the price of the next data point in the sequence. The predictor is

[1]But then, if one can predict the price, why is not everyone arbitraging the market? The problem is that, if one wants to arbitrage bitcoin prices by taking advantage of a model that predicts the price in the next 5 minutes, one would need a prediction that is essentially perfect.
[2]We always train with a random 20% of the sample reserved for cross validation.

trained to minimize both the root mean square (RMSE) and the mean absolute error (MAE).

As we briefly comment in last section, we also scale the values that feed our network, as it is known to accelerate the learning process of machine learning algorithms (as commented in [3]). Since we just need to scale the scalar value of a prediction, we use a simple MinMax function, applied over the total number of observations, as defined next:

$$X_{scaled} = \frac{X_{original} - X_{min}}{X_{max} - X_{min}} \quad (2)$$

where X_{min} and X_{max} are the minimum and maximum value of the range scaled, respectively. Clearly, this function maps the original values to the range [0,1]. With those considerations we train our model.

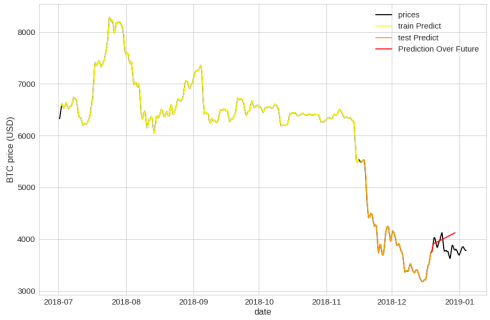


Figure 3: Network predictions after training. The black line represents the moving average measure introduced on last subsection, the yellow line the predicted value on the training part of the data, and the orange line the prediction on the test set. The part with the red line is zoomed-in and explained in Figure 4.

Figure 3 shows the quality of our regression. As we see, the model is able to predict the price of the next point with great accuracy, achieving a score of 9.91 on RMSE in training set and 10.87 on RMSE in testing set after recovering the scale.

2.3 Problems studied in this paper.

As we have seen, the idea of predicting the next point of a time series is a simple task for a neural network, and thus we prefer to focus on two more difficult problems.

The first problem we study is to predict the price of bitcoin in the next day (that is, to predict the price after 288 data points). This comes as a natural extension of the regression we have just adjusted, and it is both realistic and useful. In the following we refer to this problem as the *Next-day price*.

The second problem we study has more to do with our intention of using external data to predict prices. The motivation comes from Figure 3. If we see the orange line, every orange point takes as input the moving window of the latest 288 real data points. But what would happen if we start feeding the network with its own predicted prices? How well can we simulate the moving curves of bitcoin prices? We refer to this problem as the *Price Simulation* problem.

To understand the challenges behind Price Simulation, Figure 4 shows the poor performance of the model trained for Figure 3 when we start feeding the predicted data. Clearly, one expects that this price simulation cannot be carried away with price data only, as the high-raises and low-drops of bitcoin are probably related to external phenomena. This is where our analysis of twitter data comes in handy; we provide more details in Section 4.

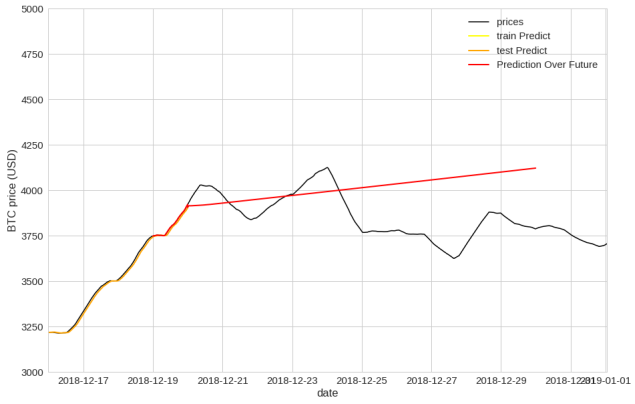


Figure 4: Simulation of Bitcoin Price. As the red curve shows, once we start feeding the network's own prediction as data we immediately loose generality and prediction power, and we get quickly stuck in an average value.

3 NEXT-DAY AND PRICE SIMULATION USING PRICE DATA

In this section we construct models that solve both the next-day price and the price simulation problem, and train them using price data only. More specifically, we want to answer the following two questions. First, can we get the same good results for next-day price as we got when predicting the next data point in the price history? The second question has to do with price simulation. As we have mentioned, in this setting we do not expect good results on the simulation, but the question remains: how far we can push this simulation using only price data?

3.1 Architecture

Let us briefly describe the architecture of our network. Choosing a recurrent network makes sense since we want to predict a single value from a time series. Moreover, LSTM networks are known for their ability to exploit temporal connections and to solve the gradient vanishing problem. Figure 5 depicts the main components of our model.

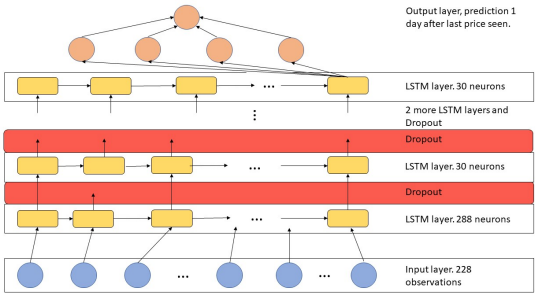


Figure 5: The architecture of our model is essentially a composition of multiple LSTM layers.

Figure 6: Next-day price prediction when all data is scaled as in Section 2. The curve is markedly displaced to an average value, not being able to recover the shape of the prediction curve and being flattened on all its extension

Our network consists of 5 LSTM recurrent layers, each consisting of 30 neurons, except from the first one, with 288 neurons. We drew these numbers from our preliminary testing, showing that such depth and width was good enough for our purposes. The input layer receives the 288 data points representing a complete day, encoding the sequential data. Then the output is passed across a sigmoid activation function and returned to feed the next layer. In these stages, the data is fed to the neural network and trained for prediction, initialized with random biases and weights. Every LSTM layer is followed by a Dropout regularization to avoid overfitting; the first layer has the biggest dropout probability and it decreases with each subsequent layer. After the last Dropout, two fully connected layers are added to obtain, first, new latent factors, and then to obtain the value that represents the prediction of the price on next day using a linear activation. When the output value is generated, is compared with the target value. The differences is then minimized using back propagation.

3.2 On scaling the data

With the introduced model, we proceed just as in Section 2, and start training our network for this new task of predicting the next day price. But the results, as Figure 6 shows were surprisingly worse: the model is practically useless.

The reason for this setback is that the prices of 1 day in the future have a much higher variation than what could be seen in consecutive data points. Together with our scaling function, this means that the network could not properly distinguish these changes. The issue is the following. If we scale all data points first, then most probably all data points of a 24 hours window would be mapped to very similar values. This would not be an issue for predicting values in the same range, but when forced to predict values for the next day, which are mapped to other ranges by the scaling, the network was no longer working how it should.

To solve this problem we use instead a scaling function that scales every 24 hours window separately, i.e., each observation of 288 points that feed our network is scaled with a common Min/Max scale function. Of course, there is a new problem with this solution: if the prices in a day vary in an (a, b) interval and the correct price in the next day is outside this interval, what should we expect the network to predict? To cope with this problem we alter the Min/Max bounds so that the lowest point (x_{min}) fed into the network does not get mapped to value 0 but to his equivalent value in scale, as a distance from the new lower bound in dataset $0.85 \times X_{min}$ and likewise for the higher point $1.20 \times X_{max}$. Figure 7 shows our results.

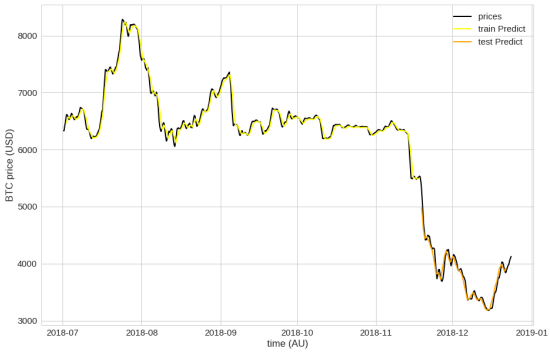


Figure 7: Next-day price prediction when all data is scaled in a per-day basis. Results are much stronger than what was obtained with a general scaling function.

There are of course many ways of improving these results: One can consider more complex networks, as in [4] where the authors develop generative adversarial networks to create a model of two part, one called Generator, who is an adversary that generate samples from time series and tries to compete with a Discriminator who

determinate if the samples are real or generated by the adversary. Others, perhaps like [12] implement attention models to be aware of particular features on the sequence, and there are several other scaling functions to consider. However, since our goal is to later combine these models with twitter data, we choose to move onto the next problem.

3.3 Price Simulation using per-day scaling

Since we have seen that per-day scaling gives us a better model for the next-day problem, it also makes sense to update our price-simulation network so that the training is likewise done with this other scaling. Results are in Figure 8.

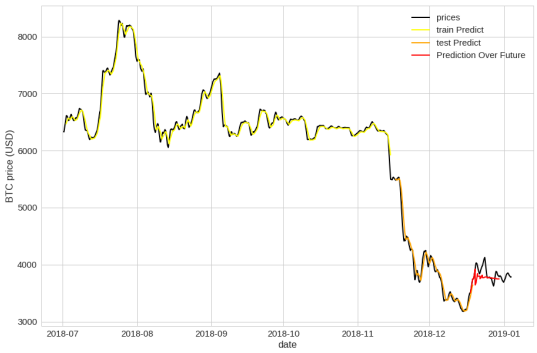


Figure 8: Price Simulation with per-day scaling. Test data is depicted on the red line

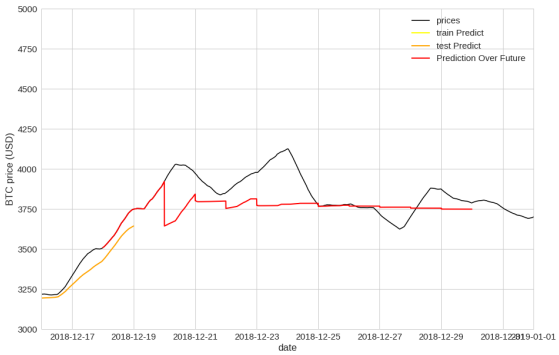


Figure 9: Zoom of the model's performance with test data. This model now manages to predict a decrease in price, but the prediction is still too coarse.

As we see, the new scaling gives our model the ability to identify slight trends over price data (shown in the red line in Figure 8). But we are still not able to detect the subsequent rise in price by the end of the test period: at this point the model has detected that prices tend to drop over time, and, unless we provide another form of external stimulus to the model, it will never change its direction. Our hypothesis is that twitter data may well be the external stimulus that we are looking for, and that "good" or "bad" tweets will give the network enough information to change the trend of what it is predicting. This is the focus of the next section.

4 INCORPORATING SENTIMENT ANALYSIS

The idea that the price of assets is affected by certain influential spokespersons is not new, and has already been studied when analyzing prices of other types of assets (see e.g. [1, 2], and our belief that twitter data is a good external source that should help in predicting Bitcoin data comes from gathering opinions of people working in the cryptomarket itself [23]. The goal of this section is to try out this hypothesis. To that extent, we first explain the way in which we collect data from these so-called influencers, then explain how to use their tweets to mine beliefs about the market, and then incorporate it into our predictor.

4.1 Twitter opinion data

There is a large body of work on how to compute the most central nodes in a social network, and to the best of our knowledge there is no agreement on a single method that fits all uses. Moreover, most of well known centrality measures (pagerank, betweeness centrality, degree centrality, etc.) require full access to the network to be computed, which is not feasible in the case of twitter. Instead, we use the following algorithm as an approximation for computing users with higher pagerank that are related to cryptocurrencies. This algorithm is inspired by [21]:

(1) We start with a seed of 10 users that are known to be influential.
(2) In each iteration, we create a list of users corresponding to all the people followed by the users in the seed.
(3) We order the list with the most followed users first, take the 10 most followed that are not already in the seed, and add them to the seed.

We selected a total of 135 influential users after 10 iterations of the algorithm. Then, we extract all tweets by these users between July and December 2018, but only store those tweets referring to one of the following terms: *bitcoin*, *btc* or *cryptocurrency* as well as their grammatical variations. This method gave us a total of 9146 tweets, which represent an average of 50.8 tweets per day across the six months considered. While this number is not small at all, is not comparable with the 288 measures per day that we use for studying price. For this reason, we propose to use a *moment*-like measure, i.e. we are going to sum all opinions over a movable window of 288 points, i. e.,

$$SM = \sum_{i}^{N} s_i \tag{3}$$

where s_i is the sentiment measure for an specific date i, while N is our movable window. We expect that tweets enclosed within this window will be a good information resource about the general appreciation of the cryptocurrency and how this is changing. Since

we are interested on extracting a measure of the meaning of each tweet, a good tool to handle this is sentiment analysis.

Note that we could have obtained more tweets by simply enlarging our list of crypto influencers. However, there is a danger in doing this: with more users in our list, the tweets by the real influential people would by diluted by tweets from regular people, which we do not trust to affect the price in the same way. One direction we are currently following is to feed both the information of the tweets and the influence of the user into the neural network, but results are still too preliminary to report in this paper.

Sentiment Analysis. Sentiment analysis is another task that has received a tremendous amount of attention over the past decade (see e.g. [16–18]). Unfortunately, we cannot simply use known tools for sentiment analysis (or even sentiment of tweets) and expect that they would transfer well to the world of bitcoin. The main reason for this is that a big number of bitcoin-related tweets are very neutral in sentiment, but convey specific information that can be very good or bad for the price. Imagine, for instance, a tweet stating *Several cities in China forbid bitcoin mining*. This tweet is very neutral in sentiment, yet conveys information that should affect the price in a negative way.

For the scope of this paper, however, we will stick to traditional methods, and see whether they can still give us something about the price, and leave specific bitcoin sentiment analysis for future work. Specifically, we use *SentiWordNet*'s sentiword tool [20], an algorithm capable of exploring the semantic relations of words using a pre-trained dictionary with positive, negative and neutral scores over certain relations. Using this tool we get a sentiment moment distribution as presented Figure 10.

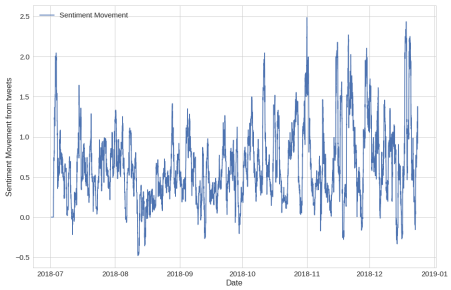


Figure 10: Evolution of sentiment movement, i.e. the sum of sentiment contained on a 1-day moving window.

It is remarkable how the sentiment score along the analyzed months is more positive than negative, with only just a few drops below zero. It is not evident if those scores are result of a generalized opinion about the topic or is just result of a high density of tweets over a particular date. However, we trust that a high density should be result of a determined event that produces a certain general opinion.

4.2 Results on prediction

With the introduction of a the sentiment measure we implement a final model, this one is similar to the described Figure 5 with the difference that the input layer is feed with observations of 288*2 dimensions, one additional dimension per point data for the moment sentiment of tweets in that point in time. With this model we train again our neural network. Our predictions are presented Figure 11 with a closer view over Figure 12.

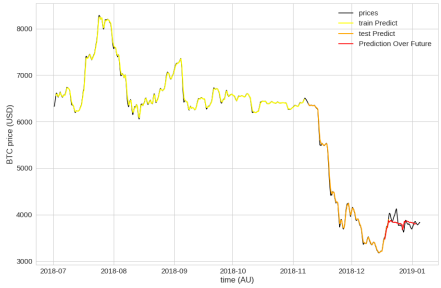


Figure 11: Prediction of the model when trained also with twitter data. The predictions for the next-day price are on the same level as the model trained without twitter data.

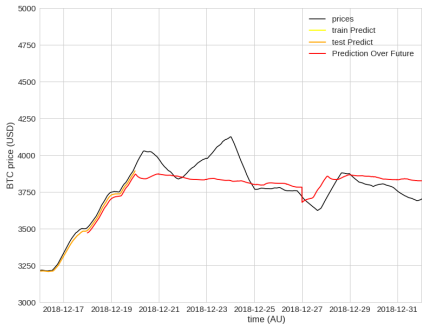


Figure 12: Zoom to final predictions. We note a fall over the seventh day that is recovered by the prediction.

Regarding the next-day problem, we noticed essentially the same predictive power, when compared via any of the two measures we were looking. However, the most interesting comparison comes from the price simulation problem, when looking at both Figure 12 and comparing it with Figure 9. We understand that the simulation of price predicted with the model of Figure 12 is much cleaner than the other one, and it is very interesting to note how twitter data predicts the movement of the curve around the 27th of December 2018. Looking at 9, we can only suspect that it was twitter data that helped shape the curve in this way.

For reference, we have included the sentiment evolution between our time window in Figure 13. We notice some trends by the end of December, but clearly there is not a huge visual correlation. It

is surprising that this data still allows us to predict spike changes within the price simulation framework.

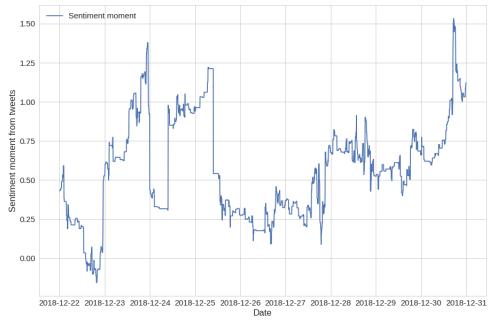


Figure 13: Evolution of sentiment movement, during prediction dates.

5 CONCLUSION AND FURTHER WORK

In studying the problem of predicting the price of bitcoin in the future, we have made advancements towards showing that the activity of influential personalities in the digital world does affect the price of this cryptocurrency.

The problem which is most interesting to us in this respect was the idea of simulating the price using a predictor model that is trained with real price data, but feeding with its own prediction and only uses twitter as its external stimulus. However, we have also shown a possible architecture that goes a long way into predicting the price when fed only with past price data: the error for predicting the next price point was almost negligible, and the predictions for the next-day price were accurate as well.

When fed with both price and twitter data, the resulting model had essentially the same predictive score than the model fitted only with price data. However, as we mentioned in the last section, the differences shown when simulating price leads us to believe that there is promising work to be done in refining our ideas.

Further work. As mentioned, evidence suggests that there is much room for improvement. In particular, we are currently pursuing the following main directions.

The first and most imperative refinement is to leverage state-of-the-art tools to produce a better classification of news into positive/negative. The main difficulty here is that the fact that a news text or an informative tweet ends up being positive or negative may not have anything in common with the actual sentiment of the tweet, as we expect most news to deliver information in a neutral or slightly positive tone. A new network classifying this precise task could be trained from scratch, but the problem here is that it is not trivial to amass the necessary data to produce a well-functioning classifier. Hence, the most promising approach appears to take advantage of the best models for standard sentiment analysis and then tune them to our specific task via transfer learning, as in e.g. [17].

The other important refinement is to study different neural architectures and different data input options for our predictors. As for data input options, we currently don't know a better way of integrating twitter data to the price stream, but we are not certain that these better ways not exists either, and perhaps there is a better way of aggregating this data. There are also several other ways of obtaining both influential users and/or their tweets: right now we are thinking on feeding the network with a greater density of tweets, but also with the information about the approximate centrality of the users posting these tweets so that the non-influential tweets do not end up diluting the valuable information. We are also interested on adding other types of data into the mix, such as news and reddit posts (again posted by influential users).

As for network architectures, one of our last ideas is to recover the results of Zhou et al. (2018) [4] about generative adversarial nets. The intuition here is that an adversarial net would use price history to train next data points, but an adversary would then use external twitter data to discriminate against the likelihood of the predicted price. This would perhaps not only improve our predictions, but would also give us a much more cleaner justification on the relation between price and tweets. Other ideas have to do with adding further penalization to the loss function so that models correctly predict inflection points.

REFERENCES

[1] Xi Zhanga Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fanga, Philip S. Yu. (2018). Improving Stock Market Prediction via Heterogeneous Information Fusion.
[2] Xi Zhang, Yixuan Li, Senzhang Wang, Binxing Fang, Philip S. Yu. (2018). Enhancing Stock Market Prediction with Extended Coupled Hidden Markov Model over Multi-Sourced Data.
[3] Ting Li, Bingyi Jing, Ningchen Ying , Xianshi Yu. Adaptive Scaling (2017).
[4] Xingyu Zhou, Zhisong Pan, Guyu Hu, Siqi Tang, and Cheng Zhao. (2018). Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets.
[5] Kasun Bandara, Peibei Shi, Christoph Bergmeir, Hansika Hewamalage, Quoc Tran, Brian Seaman. (2019). Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology.
[6] Zahra Karevan, Johan A. K. Suykens. (2018). Spatio-temporal Stacked LSTM for Temperature Prediction in Weather Forecasting.
[7] Uma M. Girkar, Ryo Uchimido, Li-wei H. Lehman, Peter Szolovits, Leo Celi, Wei-Hung Weng. (2018). Predicting Blood Pressure Response to Fluid Bolus Therapy Using Attention-Based Neural Networks for Clinical Interpretability.
[8] Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, Layla El Asri. (2018). From FiLM to Video: Multi-turn Question Answering with Multi-modal Context.
[9] Hyunwoo Jung, Moonsu Han, Minki Kang, Sungju Hwang. (2018). Learning What to Remember: Long-term Episodic Memory Networks for Learning from Streaming Data.
[10] Vidya Prasad K, Akarsh S, Vinayakumar R, Soman KP (2019). A Deep Learning Approach for Similar Languages, Varieties and Dialects.
[11] Bilal Wehbe, Octavio Arriaga, Mario Michael Krell, Frank Kirchner. (2018). Learning of Multi-Context Models for Autonomous Underwater Vehicles.
[12] Guoqiang Zhong, Xin Lin, Kang Chen. (2018). Long Short-Term Attention.
[13] Dzmitry Bahdanau, Dmitriy Serdyuk, Philémon Brakel, Nan Rosemary Ke, Jan Chorowski, Aaron Courville, Yoshua Bengio. (2016). Task Loss Estimation for Sequence Prediction.
[14] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, Guy Van den Broeck. (2018). A Semantic Loss Function for Deep Learning with Symbolic Knowledge.
[15] Sean Welleck, Zixin Yao, Yu Gai, Jialin Mao, Zheng Zhang, Kyunghyun Cho. (2017). Loss Functions for Multiset Prediction.
[16] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. (2018). Deep contextualized word representations.
[17] Jeremy Howard, Sebastian Ruder. (2018). Universal Language Model Fine-tuning for Text Classification.
[18] Shaunak Joshi, Deepali Deshpande. (2018). Twitter Sentiment Analysis System.
[19] Sara Rosenthal, Noura Farra, Preslav Nakov. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter.
[20] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. (2006). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.

[21] Daniel Pérez. (2018). Crypto-Influencers Algorithm. Url = “https://github.com/dperezrada/crypto-influencers”, lastaccessed on June 23, 2018.
[22] Nakamoto, Satoshi. (2018). Bitcoin: A peer-to-peer electronic cash system.
[23] Daniel Pérez Rada (2018-01-22). “Personal communication”.
[24] Joost van der Burgt. (2018). Making Sense of Bitcoin Price Levels.

Recommender Systems for Online Video Game Platforms: the Case of STEAM

Germán Cheuque
IMFD &
Pontificia Universidad Católica
Santiago, Chile
gacheuque@uc.cl

José Guzmán
Pontificia Universidad Católica
Santiago, Chile
jaguzman6@uc.cl

Denis Parra
IMFD &
Pontificia Universidad Católica
Santiago, Chile
dparra@ing.puc.cl

ABSTRACT

The world of video games has changed considerably over the recent years. Its diversification has dramatically increased the number of users engaged in online communities of this entertainment area, and consequently, the number and types of games available. This context of information overload underpins the development of recommender systems that could leverage the information that the video game platforms collect, hence following the trend of new games coming out every year. In this work we test the potential of state-of-the-art recommender models based respectively on Factorization Machines (FM), deep neural networks (DeepNN) and one derived from the mixture of both (DeepFM), chosen for their potential of receiving multiple inputs as well as different types of input variables. We evaluate our results measuring the ranking accuracy of the recommendation and the diversity/novelty of a recommendation list. All the algorithms achieve better results than a baseline based on implicit feedback (Alternating Least Squares model). The best performing algorithm is DeepNN, the high order interactions are more important than the low order ones for this recommendation task. We also analyze the effect of the sentiment extracted directly from game reviews, and find that it is not as relevant for recommendation as one might expect. We are the first in studying the aforementioned recommender systems over the context of online video game platforms, reporting novel results which could be used as baseline in future works.

KEYWORDS

Recommender System, Factorization Machines, Deep Neural Networks, Deep Factorization Machines, Novelty, Diversity

ACM Reference Format:
Germán Cheuque, José Guzmán, and Denis Parra. 2019. Recommender Systems for Online Video Game Platforms: the Case of STEAM. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316457

1 INTRODUCTION

According to the European Mobile Game Market, in 2016 over 2.5 billion people spent part of their time playing video games. The huge number of adepts have made the video game industry one of the most valued in the world. In 2017 alone this industry grew by 10.7% in earnings, achieving over $116 billion. One of the biggest reasons behind this success is the high diversification that this industry has had over latest years. We can now find multiple platforms dedicated to games, an enormous number of genres and game' categories and even platforms that provide social interaction between players. This adaptability has generated a huge number of items with different attributes that are able to attract diverse users. An example of adaptation is the STEAM[1] platform, a company dedicated to the digital distribution of video games which allows users to buy, see, share opinions, play on-line, play off-line, compete and cooperate across the platform and games. Over 10 million people connect to the STEAM server each hour to play video games. However, the fact of having such a variety of products and so many users, makes it difficult to choose a particular new game some user like. Also, according to STEAM registries of 2014, about 37% of games purchased have never been played by the users who bought them. This context creates a need for recommender systems, which are systems able to make relevant personalized suggestions [8], alerting users to unknown games as well as new releases.

Playing video games is a recurrent activity, in this sense that is it more similar to music listening than to movie watching. A preferred game is played many times, but users also want to discover new games. This represents a double challenge for the industry: the need for video games which encourage users to return, as well as helping users find novel games which will be consumed as much as those already liked. Our intuition is that the great number of features available in the STEAM platform allows us to explore interactions between items' features, as well as between users and their preferences. These relations make it reasonable to argue that it is possible to face the second challenge. Being able to develop an algorithm which addresses the aforementioned challenges could generate great benefits for the industry, the community of users and even for game developers, by predicting what the users want the most as well as promoting new releases.

To achieve these goals, we test three state-of-the-art algorithms based on different recommendation paradigms that can be used with different types of input data:

- Recommender System (RecSys) based on Collaborative Filtering (CF) that uses the ALS (Alternating Least Squares) algorithm for making recommendation based on implicit feedback [4]. We use it as a baseline.

[1] https://store.steampowered.com/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316457

- RecSys based on Factorization Machines [10] that benefits from a huge amount of data, context and features from users, items and purchases.
- RecSys based on FM and deep learning techniques (DeepNN and DeepFM [3]), where we introduce sentiment analysis to decode information contained on reviews.

We argue that feeding our models solely with the item preference history of users (implicit feedback [4, 7]) should be enough to make a good baseline recommendation model. Moreover, we think that adding content and contextual information could be effectively used by the FM and DeepNN models to improve ranking and novelty performance performance, due to interaction effects. However, the combination of both in a joint model (DeepFM) could either produce an improvement or rather be just redundant with respect to DeepNN.

Our contributions are the following: (i) We are the first in studying the aforementioned state-of-the-art recommender systems on the context of video games, reporting novel results to be used as baselines in future works, (ii) in addition to ranking accuracy, we report and analyze results about novelty and diversity of recommendations, which is critical in the online video game industry, and (iii) we are the first in analyzing the impact of using sentiment analysis of the textual reviews over the quality, novelty and diversity of recommendations of online video games.

This work is organized as follows: Section 2 presents a discussion of the state of the art, the studies that inspired our work. In section 3 we present the Materials (datasets) and the methods (recommendation models) used in this work. Section 4 shows the experimental methodology and evaluation metrics to measure the performance of our systems, while in section 5 we present a sensitivity analysis of parameters. Section 6 shows our most relevant results and finally, section 7 discusses our conclusions and ideas for future work.

2 RELATED WORK

Numerous are the studies that address the recommendation problem, however just a few have emerged in the video game industry. The video game purchase platforms or in the video game communities an interesting places to develop and test the potential of existing recommender systems. Despite this, there are some notable studies whose application could be relevant for the introduced context.

Bertens et al. 2018 [1], for example, tries to measure the performance of two recommender systems based on the idea of prediction of the next most likely item to buy. The paper works over the behavior of Japaneses card players of the game *Age of Ishtaria* taking into consideration their play time, daily progress and purchase history. They test two models: the first one is called Extremely Randomized Trees (ERT), which is a randomized version of the decision tree algorithm. This method has the advantage of being computationally efficient due to its high parallelization ability, and it also has the property of preventing overfitting. The second one corresponds to an algorithm based on deep neural networks, with the particularity of being built as a recurrent network because of the sequential property of data. Both algorithms show very similar results, with ERT being the one with better performance in both train time and scalability.

Quadrana et al. 2018 [9] propose a recommender system based on attention to a sequence of events as one of the best approximations to solve the recommendation task on the introduced context. They survey multiple developed models, considering different tasks and goals to achieve. On their discussion, they highlight the potential of neural networks for the recommendation task. Although their discussion is relevant to the video game industry context, they emphasize their implementations to be tested on the *in session* attention context. One relevant approximation of this type of application can be found in Wan et al. 2018 [13] where the authors study the so called *monotonically behavior chains* concept that is understood as a sequence of events that account for a more explicit preference that a user has about an item. This chain is constructed from implicit and abundant information to a more explicit and limited information. The novelty of the publication lies on the capability of introducing a new paradigm to face the recommendation task and being able to prove its potential over different datasets, including one about video games from STEAM. Their improvement in performance went from 1% to 28% with respect to *Most Popular* method, depending on the dataset.

Due to the context of networks and interactions that exist in video game platforms, recommender systems based on graphs could be very useful as a modeling approach. In this topic, Shams et al. (2016) [11] is one of the most recent studies. In this publication they introduce the method GRank to correctly model the priorities of users and to discriminate better the relevant connections between nodes.

In this work, we propose to develop a recommender system for games based on state-of-the-art techniques, by comparing the results of implicit feedback collaborative filtering [4], factorization machines [10] and deep neural networks [3]. Unlike previous work, our system leverages interactions between users and items by mixing up implicit information from users and features from items, considering implicit information as playtime and explicit information like opinions, within reviews. In addition, we evaluate beyond prediction or ranking accuracy, we use the metrics proposed by Vargas and Castells [12] to measure novelty and diversity, since these metrics are critical for the continuous and diverse releases in the game industry.

3 MATERIALS AND METHODS

3.1 Materials: The Dataset

To carry out the implementation of the mentioned recommender systems (discussed in detail on the next section) and evaluate its predictive ability we work over three datasets, obtained from the collection shared by J. McAuley[2]. The first one consists of the purchase history of Australian users of the STEAM platform, a database sorted by user, indicating for each one the list of purchased items with a small collection of metadata as the playing time. The second dataset contains the opinion that different users on the platform have about the available items, this opinion serves as a review for

[2]http://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data

the whole community. The third and last dataset gives us detailed information about different features of video games, such as the genres they belong to or their platform availability (as discussed in [5]). The data is collected between October 2010 to January 2018.

The different datasets are joined with the purpose of creating tuples of data associated with each pair of item and user. Table 1 shows a summary of final descriptive statistics. Each tuple consist of 32 features where we find abundant implicit information, as the playtime, and scarce explicit information as the *recommend* feature. As we can see, there is no general rating measure that could be used as a measure of how much a user likes or prefers an item. Because of this, our analysis and the model we consider is based on measures of implicit feedback or made from a combination of information that we consider useful in making recommendations. The wealth of information gives us a unique chance of testing different kinds of data and their explicit and implicit interaction, as well as their non-linearity.

Special attention is given to *playtime*, considering the importance of this variable reported in previous research on recommender systems [7, 14].

Table 1: Summary of features.

Feature	Type	Description
user_id	Str	Unique identifier of a user
item_id	Str	Unique identifier of item
count	Int	Number of games purchased by the user
playtime	Int	Period of time that the game has been played by a user, in hours
RecCount	Int	Times that an item has been recommended
Metacritic	Int	Valuation of game, agreed to specialized critic.
GenreIs |G| = 13	[Bool]	True if the item belongs to one of the 13 different game genres.
Category |C| = 8	[Bool]	True if the game belongs to one of the 8 game categories, e.g.: multi player
Platform |P| = 3	[Bool]	True if the game is supported by one of the 3 listed operative systems.
recommend	[Bool]	True if user recommends the item.
review	Str	Free text which reports the user opinion.

About 5, 153, 209 of tuples are contained in our dataset with a total of 70, 912 different users and 10, 978 different items (video games) to choose from. These numbers tell us about a large sparsity of 0.66% seen registries. A large sparsity is not ideal for training a model of interactions between users and items. Because of this, we filter the dataset; we believe that densities in the order of 10% could be useful for our objectives. Our final dataset considers only items purchased at least 200 times and users with at least 100 purchased items. A summary of our final database is presented in Table 2. With the data filtering, we obtained a new dataset with a density close to 10%, reducing its original size to a half, but still of a considerable size to carry out experiments. We also analyze the distribution

Table 2: Summary of statistics of the final database after cut.

Descriptive statistic	
Number of registries	2,149,858
Number of different users	8,183
Number of different items	2,872
Number of reviews	9,823
Dataset density	9.14%
Average purchases per user	262.72
Average purchases per item	748.55
Average hours played per user	161,317.9
Average hours played per item	476,727.9

of consumption in the dataset, in order to dismiss potential issues due to imbalance bias: too few users accounting for too many of the total transactions. Figure 1 shows that user bias is small: around 67% of them explain the 80% of tuples. However, Figure 2 shows that a small number of items (40%) are responsible for most records (80%). This situation encourages to further analyze our systems to ensure they recommend different kinds of items. To address this concern, we are going to use novelty and diversity metrics. Further details are discussed in section 4.

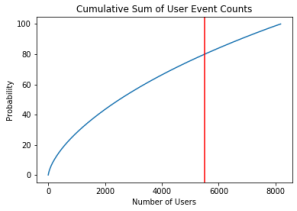


Figure 1: Cumulative sum of contribution of users to the number of registries. Sixty-seven percent of users explain the 80% of data.

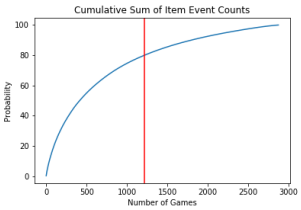


Figure 2: Cumulative sum of contribution of items to the number of registries. Forty percent of items explain the 80% of data.

Implicit feedback data. We will work on implicit feedback to learn user preferences, specifically, using playtime as our proxy measure. After preliminary analysis of the whole playtime distribution (see Figure 3), we proposed a limit of 5 hours as enough to differentiate between preference or not. This is indeed a strong

assumption, since the ideal would be knowing the usual playtime of each game as a more precise measure. But for the sake of simplicity, we leave this idea for future work. Figures 3 and 4 show distributions of playtime for the most active users and most purchased games.

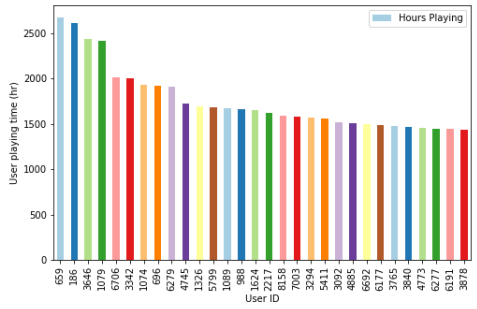


Figure 3: Playtime of most active users.

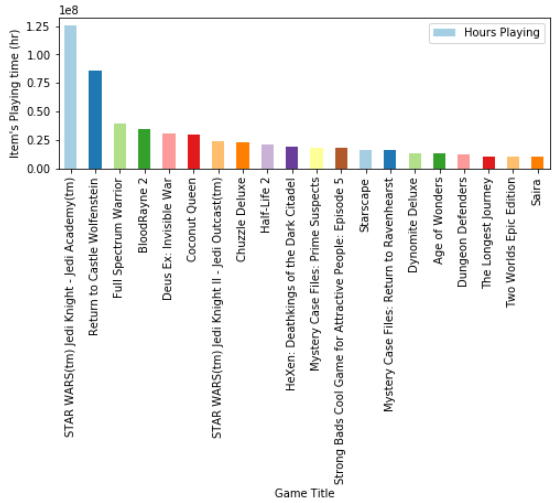


Figure 4: Playtime of most purchased games.

Figure 3 shows small differences on playtime among the most active users playing video games on STEAM. Meanwhile, most purchased games show bigger variations on playtime, where we find games widely played across users and also, games not played at all. When we see the average playtimes in table 2 we can infer another characteristic about the distribution of user and games. About 161,317.9 hours on average are played per user, while a total a 476,727.85 hours is the average time that an item is played on record history. Again we find that users have a more homogeneous distribution compared to items, for which a rather small fraction of items represent most interactions in the dataset. Although these numbers could suggest us that a low potential for making both accurate and diverse recommendation lists, when we apply over our records a threshold of 5 hours playtime, we find that 46% of database are over this threshold and, therefore, correspond to relevant items for recommendation task.

3.2 Methods: Recommender Systems Models

3.2.1 *Alternating Least Squares (ALS).*.

Under the diversity of models of matrix factorization, the ALS model stands out for being capable of working using implicit feedback [4].

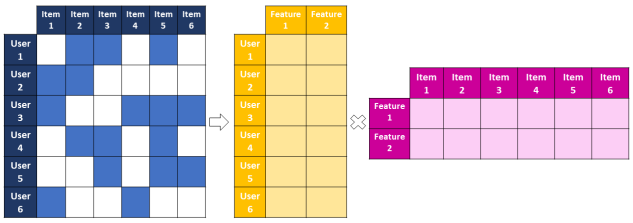


Figure 5: Matrix factorization model representation.

Figure 5 shows the idea under the factorization method, This is, to find a new representation for user (x_u) and items (y_i) interactions, so that a particular preference of an user about an item is given by the dot product of the latent factor representation of each of them. An innovation that introduces this algorithm analyzes the implicit feedback measure through the insertion of two new measures that represent the preference and confidence that exist behind the uptake of an item. Preference tells us if an item is consumed or not.

$$p_{ui} = \begin{cases} 1 & \text{if} r_{ui} > 0 \\ 0 & \text{if} r_{ui} = 0 \end{cases} \quad (1)$$

Where r_{ui} is some measure of implicit feedback. The second measure stands for the confidence of that preference,

$$c_{ui} = 1 + \alpha * r_{ui} \quad (1)$$

Where α is a linear scale factor that set more importance to relevant items, above never played ones. The value $\alpha = 40$ is usually used as a result of the original paper [4]. Under these definitions, the search of latent factor for users and items is made by the optimization of the following loss function,

$$min_{y*} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_u ||x_u||^2 + \sum_i ||y_i||^2), \quad (2)$$

the optimization via least squares that give its name to the algorithm is visible. In this way, subsequent updates of latent factors are given by,

$$x_u = (V^T V + V^T(C^u - I)V + \lambda I)^{-1} V^T C^u p(u) \quad (3)$$

$$y_i = (U^T U + U^T(C^i - I)U + \lambda I)^{-1} U^T C^i p(i) \quad (4)$$

Some relevant parameters like the number of latent factors to use are selected through analysis in many iterations. In this article, we

used the ALS implementation found in the pyreclab[3] recommender library.

3.2.2 *Factorization Machines (FM)* .

Factorization Machines (FMs) are a type of model for recommender systems making latent factor models as easy to use as regression or SVM models [10]. FMs can deal with different types of inputs, from continuous to discrete variables, and more importantly, latent factor models. FMs can model interactions of different orders (order-n) between these variables [10]. In our case, $n = 2$ was defined because it has been widely used in this way as it delivers good results while maintaining a reasonable training time. With this configuration, the output is a prediction resulting from the linear interaction of the inputs (order-1) plus that of the latent factors, defined by the relationships between pairs of inputs (order-2), as observed in the equation 5.

$$y(x) := w_0 + \sum_{i=0}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \qquad (5)$$

We decided to choose this model because it is a type of system that can scale quickly, has shown excellent performance when working with sparse datasets, and supports numerical input data of any type. These are all features that contribute to the recommendation process in a company like video games, where product variety and growth are critical factors in the process.

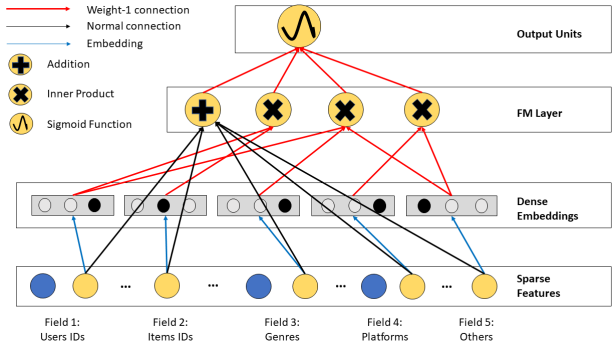


Figure 6: Factorization Machine Architecture. Image adapted from [3].

Figure 6 (as in [3]) shows an example of the architecture of a factorization machine model and the Figure 7 (as in [10]) an example of the inputs and outputs that can receive.

3.2.3 *DeepFM*.

DeepFM is a newer model that seeks to take advantage of the versatility of Factorization Machines to model low-order interactions between input variables, with the ability to model deeper interactions of Deep Neural Networks (DNN) [3]. To do this, a DNN is implemented in parallel to the FM and the result of both is joined

[3]https://github.com/gasevi/pyreclab

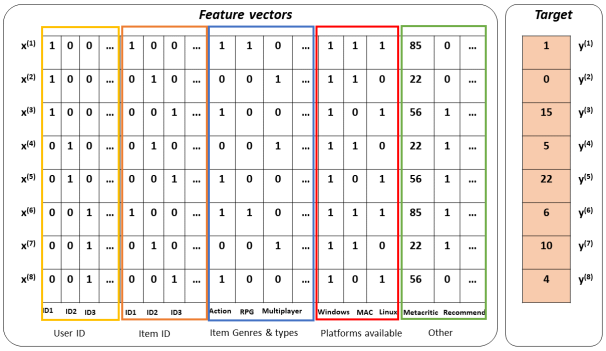


Figure 7: Input example of features received by FM. Image adapted from [10].

on a last node using a sigmoid function, making the output a prediction resulting from the interaction of several orders of inputs that represents the probability of whether or not a user belongs to a particular class defined with training tags. In our case, this class referred to whether the user liked the game or not.

We decided to classify the entries according to their type (user, item, category, platform, etc.) and to use an embedding of the entries that gives the same weight to each group, so that the effect of the "sparse" variables did not interfere in a significant way against the effect of the other variables. This embedding consisted in generating several matrices that allowed transforming the N_i variables of each input class into a fixed K number defined as a model parameter. This idea was taken from the same paper [3] from where the model information was obtained, providing us with a large part of its implementation [4], which, for the most part, is made using the tensorflow library. Figure 8 (as in [3]) shows a representation of all the components of the model and how they are connected.

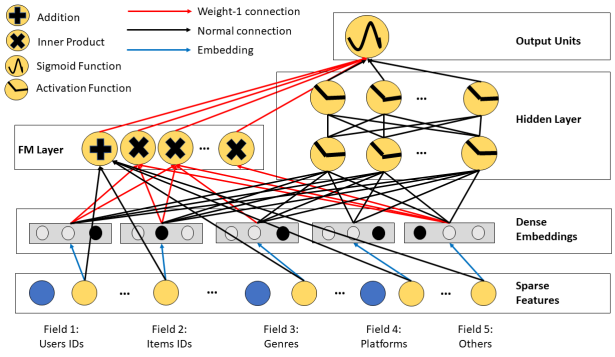


Figure 8: Example of the DeepFM Architecture. Image adapted from [3].

3.2.4 *DeepNN*.

The novel component of DeepFM is the deep neural network (also

[4]https://github.com/ChenglongChen/tensorflow-DeepFM

referred to by us as DeepNN) which analyzes in parallel the interactions between users and items. The deep component is a feed-forward neural network and is used to learn high-order interactions. A summary of this architecture is shown in Figure 9.

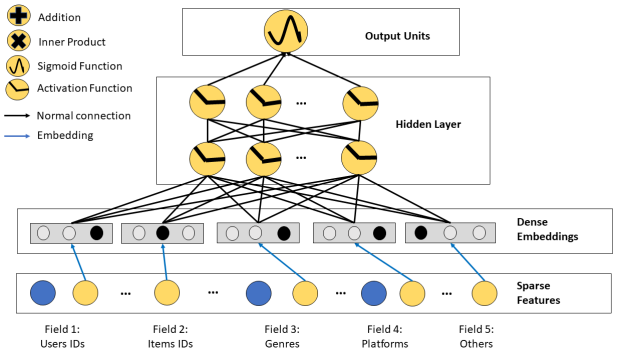


Figure 9: Example of the DeepNN component of the DeepFM Architecture. Image adapted from [3].

Its first component is an embedding layer that compresses the input fields to a low dimensional and dense real-valued vector. Notice that these fields can be of different lengths while embedding always returns a codification of same K units of length. We denote the output of the embedding layer as:

$$a(0) = [e_1, e_2, ..., e_m], \tag{6}$$

where e_i is the i-th field embedded from a total of m fields. These inputs are fed then to the hidden layers where the features interact. The forward process is:

$$a^{(l+1)} = \sigma(W^{(l)}a^{(l)} + b^{(l)}), \tag{7}$$

where l is the layer depth and σ is an activation function. The variables $a^{(l)}, W^{(l)}, b^{(l)}$ are the output, model weight, and bias of the l-th layer. Then a last dense vector is generated and these results are passed throughout a sigmoid function that returns a final output for the recommendation task.

3.3 Sentiment analysis

One of the most interesting features in our dataset corresponds to reviews. A review is a piece of text that encloses the opinion that a user has and gave. However, a piece of text is not useful itself to feed our models. To solve this, we propose sentiment analysis techniques to transform this opinion to a numeric measure that can give us a representation about how positive or negative the expressed idea is. Using it could help the confidence degree obtained from the explicit *recommend* feature, being also able to impute it when those values do not match. We experiment with three different tools.

First we use the open source algorithm *Tweetment*. This algorithm exists as a Python library and is capable of making binary sentiment classification to labels positively or negatively, represented by values 1 or 0, respectively. This algorithm is trained over a tweet database, achieving a F1-score of 69.02% on binary classification [6]. Another consulted algorithm was *SentiWordNet*, developed as an opinion mining application able to classify text in a continuous range [-1, 1], indicative of how positive or negative is the expressed idea. This algorithm works using a pre-trained dictionary that synthesize n-grams relations, giving a score about its positiveness, negativeness and objectiveness, without being exclusive [2]. Finally, we tried a model base on Neural networks and a word2vec embedding. This algorithm tries to pay attention on text over the game in question. However encountered two drawbacks in its development: we were able to make just binary classification and also, the implementation achieved a F1-score lower than Tweetment. We finally decided to work with SentiWordNet because of the advantage of continuous measure for the sentiment appreciation. This could be more appropriate for different comment context (jokes, etc).

4 EXPERIMENTAL METHODOLOGY

The experiments presented the following sections consist of the training and evaluation of the recommendation models. To ensure robustness we proceed to make randomized folds of the dataset (k-folding), applying cross validation techniques to average these scores. To train ALS algorithm, we use five folds while on FM and DeepFM we use only three. The testing set was also randomly obtained, however, because of our previous cut on dataset, we have items that are purchased at least 100 times, so we verify that at least ten of these purchases stay on test set so we try to predict the user's preferences is preference about them and verify what recommendations our models suggest.

To measure our algorithms performance we use several metrics. Training time is considered as a relevant measure to give us an idea about the scalability of the models. Also, we want to measure the Mean Average Precision (MAP) at the 10th top position in a list. So, Average Precision is defined as,

$$AP = \frac{\sum_k P@k * rel(k)}{\#relevant_items} \tag{8}$$

and the Mean average precision measure is given by,

$$MAP = \frac{\sum_{u=1}^{n} AP(u)}{\#users} \tag{9}$$

Similarly we want to calculate the nDCG measure at the 10th position of the list. This measure tries to standardize the gain and utility of a recommendation list through the consideration of the relevance of items and its ranking position on the list (introducing a logarithmic discount). The definition of this parameter is a ratio between the Discounted Cumulative Gain at position k, given by:

$$DCG_k = \sum_{i=1}^{k} \frac{rel(i)}{log_2(1+i)} \tag{10}$$

and its ideal value at position k, also known as Ideal DCG. So the nDCG measure is given by,

$$NDCG_k = \frac{DCG_k}{IDCG_k} \tag{11}$$

Finally we are interested in measuring how diverse and new are the recommendation made by our systems. As we have already discussed, a large number of new games are released each year, making these metrics very relevant. To measure these quantities we use similarity based relations between items from a recommendation list. Details are presented in [12].

$$Novelty(R|u) = \sum_{n,j \in u} disc(n)p(rel|i_n, u)p(rel|j_n, u)d(i, j) \quad (12)$$

$$Diversity(R|u) = 2 \sum_{k<n} disc(n)disc(k)p(rel|i_n, u)d(i, k) \quad (13)$$

The definitions introduce multiple discount factors, used as logarithmic ones. The d(i,k) term refers to the distance between pairs of items, this was calculated as a cosine distance. Finally the terms $p(rel|i_n, u)$ refer to the probability that a user u prefers the item i_n, a factor that could be understood as a relevance. This measure is obtained directly from the output of recommender systems.

5 PARAMETERS SENSITIVITY ANALYSIS

With the purpose of obtaining the best results, we proceed to analyze the different tunable parameter. ALS algorithm has three parameters that we can change they: The number of latent factors used to represent the factorization, the number of epochs of training also refereed to as iterations; finally the λ factor, used for regularization on equations 3 and 4. The sensitivity analysis is shown on the next figures.

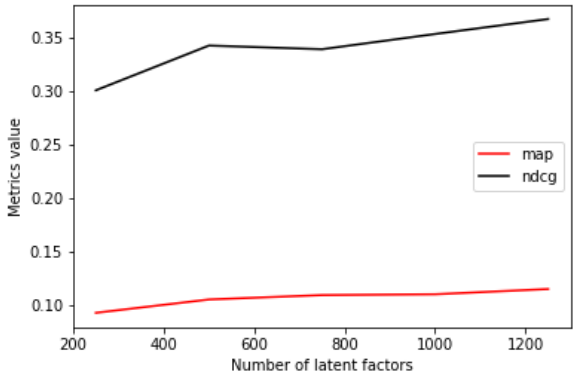


Figure 10: Training with different number of latent factors.

We finally work with the combination of the best found parameters. We set regularization to 0.01, a number of 500 latent factors privileging the training time against low improvements over 800. Also 300 iterations were used for training.

The deep learning models have two tunable parameters. The first one directly influences the net architecture, referring to the number of neurons in each layer. The second one is the batch size, that encloses the number of tuples that are analyzed together. The last one is crucial due its responsibility on the over-fitting and for being

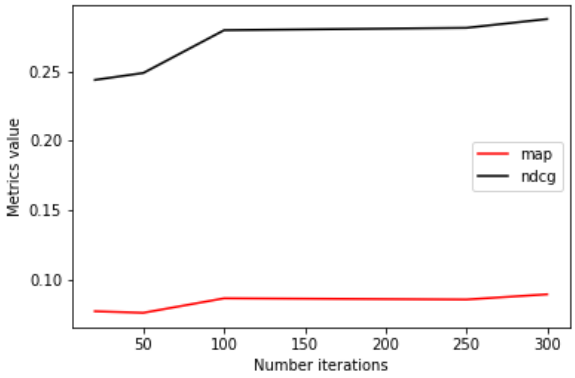


Figure 11: Training with different number of iterations.

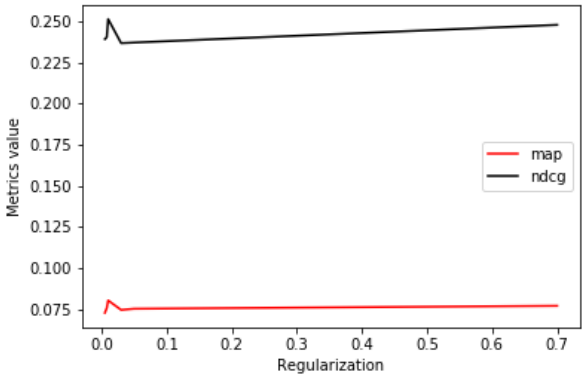


Figure 12: Training with different regularization parameter.

capable to change the learning ratio. Our training produces graphics like the ones shown in figures 13 and 14. The presented score is measured in terms of the Gini norm, a good metric for binary classification that is equal to ROC-AUC but with a bigger resolution spectrum.

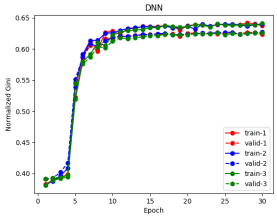


Figure 13: DNN training curve.

Using those curves and the defined metrics, we produced a sensitivity analysis shown on table 3.

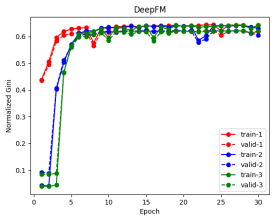


Figure 14: DeepFM training curve.

Table 3: Sensitivity analysis for DeepFM model. We vary the number of neurons per layer (H_{ixj}) and the batch size (B_n). We use as a base a default configuration with a batch size of 1024 and 32 neurons by layer.

Configuration	Time (sec)	MAP @10	NDCG @10	Novelty	Diversity
B_{1024} HL_{32x32}	2027,52	0,8911	0,9437	0,1667	0,4245
B_{512}	2186,77	0,8940	0,9456	0,1845	0,4967
B_{256}	2692,96	0,8919	0,9440	0,1847	0,4996
HL_{16x16}	1815,70	0,8939	0,9454	0,1815	0,4775
HL_{8x8}	**1708,57**	0,8941	0,9455	0,1837	0,4869
B_{512} HL_{8x8}	2180,6	**0,8945**	**0,9458**	**0,1857**	**0,5149**

We can see that when batch size is decreased from 1024 to 512, we get better results under all metrics. However, if we continue decreasing this value to 256 the results are slightly worse except for the Novelty that increased a little. That can be explained because when we decrease the batch size we generate an over-training on the training set, and relations are quickly learned with a few number of records. Also, when we decrease the number of neurons by layer from 32 to 8, we get better results over all metrics. This effect can be explained because of the number of features on each tuple (32), so when we use a smaller number of neurons we are forcing an embedding that benefits from interactions. Finally, we combine those configurations to create a model with the best performance for use in the test set.

6 RESULTS

Table 4 shows the results of evaluating all recommender systems models discussed previously.

Table 4: Test results for every model presented, including results with and without (WS) sentiment analysis considered.

Configuration	Time (seg)	MAP @10	NDCG @10	Novelty	Diversity
ALS	1421,8	0,107	0,332	-	-
FM	1450,86	0,893	0,944	0,176	0,45
DeepNN	1889,58	0,897	0,947	0,186	0,49
DeepFM	2027,52	0,891	0,943	0,167	0,43
FM(WS)	**1416,04**	0,894	0,945	0,18	0,46
DeepNN(WS)	1866,35	**0,897**	**0,948**	**0,197**	**0,54**
DeepFM(WS)	1984,17	0,892	0,944	0,19	0,53

First of all, we see that our baseline model ALS is the fastest to train, but is the one with the poorest performance in all metrics discussed. The performance of FM, DNN and DeepFM is close to eight times better in terms of MAP@10, indicating that the content and contextual data are very important for this problem, and implicit co-ocurrances are not enough to yield good results. Another important result is that contrary to our assumptions, the DeepNN model achieves better results on all metrics outperforming FM and even DeepFM that includes this architecture on it. This could indicate that the high order interactions contributed from the DeepNN model provide more information than the lower order ones from the FM, so these interactions are generating noise to the final result of DeepFM. Although the differences in terms of MAP@10 and NDCG@10 are small, they are consistently better for DeepNN. Moreover, novelty and diversity are also better for DeepNN, showing that even with similar ranking results, the resultant recommendations are more novel and diverse, which are very important aspects in the game industry.
Another interesting result is that when we do not consider the sentiment analysis values generated from reviews, all models show even better performances against themselves when considering this feature. A reasonable cause for this is the small proportion of reviews compared to the total user-game interaction records. The metrics of Novelty and Diversity also benefit from not including sentiment, reaching up to 0.54 in terms of diversity for DeepNN without using sentiment. As we have commented, on in the proposed online video game problem, it is always better to reach improvements in diversity and novelty in addition to accurate predictions as assessed with MAP and NDCG.

7 CONCLUSION

In this work we experimented with different options of recommendation systems within the context of video game recommendation, using a dataset from the company STEAM as a test platform. We joined three databases that contained information of user purchases, the hours dedicated to play each item, their social interaction (critics) and the characteristics of video games. The ALS model was chosen as the base model for comparison, and tests were carried out with the models of Factorization Machines (FM), DeepNN and DeepFM. The latter model uses a deep neural network (DeepNN) that works in parallel to a layer composed of a FM, to introduce higher-order interactions between inputs, aiming to improve the performance of our prediction in terms of factors of novelty, diversity and accuracy.

All the models studied outperformed the ALS baseline model. DeepNN stood out from the rest. Despite being a simpler model than DeepFM, it managed to exploit item-user relationships better, and although it takes longer than FM and DeepFM to achieve competitive results, it achieves consistent results over different datasets. Also, it obtains better results with the evaluated metrics, which implies that the higher order interactions provided by the DeepNN model provide more information than the lower order interactions provided by the FM model. As we argued the non-linear interaction from items

features is leveraged by deep-learning based methods, but interestingly, this difference is better perceived in novely and diversity improvements over the lists of recommendations.
In order to obtain additional information, we used the user reviews with different methods of sentiment analysis, finding a continuous measure that could further abstract the user's opinion for the video game. We conclude that the proportion of available reviews was not large enough in order to contribute to an improvement of the results, being translated instead in an effect of noise. Metrics reduce their value as much in MAP@10 as in NDCG@10 and even in the measures of Novelty and Diversity.

In future work we plan to perform tests of parameter analysis on the DNN model that gave the best results to see if they can be improved. We might also conduct a user study to validate the offline results on a real system, where other variables can have a strong impact on the final results. Finally, we have recently found that the STEAM database has been updated by its author. In its last version, a new dataset of 7,793,069 reviews is present, so we expect for future work to leverage them for a larger text analysis. This will help us test whether sentiment analysis and other textual features can indeed be a useful tool for recommending video games.

8 ACKNOWLEDGEMENTS

This work have been partially funded by Millennium Institute for Foundational Research on Data (IMFD). We have also been partially supported by NIC Chile.

REFERENCES

[1] Paul Bertens, Anna Guitart, Pei Pei Chen, and África Periáñez. 2018. A Machine-Learning Item Recommendation System for Video Games. *arXiv preprint arXiv:1806.04900* (2018).
[2] Andrea Esuli and Fabrizio Sebastiani. 2007. SentiWordNet: a high-coverage lexical resource for opinion mining. *Evaluation* 17 (2007), 1–26.
[3] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1725–1731.
[4] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 263–272.
[5] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.
[6] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*. 321–327.
[7] Denis Parra and Xavier Amatriain. 2011. Walk the talk: analyzing the relation between implicit and explicit feedback for preference elicitation. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*. Springer-Verlag, 255–268.
[8] Denis Parra and Shaghayegh Sahebi. 2013. Recommender systems: Sources of knowledge and evaluation metrics. In *Advanced techniques in web intelligence-2*. Springer, 149–175.
[9] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *ACM Comput. Surv.* 51, 4, Article 66 (July 2018), 36 pages. https://doi.org/10.1145/3190616
[10] Steffen Rendle. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 57.
[11] Bita Shams and Saman Haratizadeh. 2017. Graph-based collaborative ranking. *Expert Systems with Applications* 67 (2017), 59–70.
[12] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 109–116.
[13] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 86–94.
[14] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 113–120.

The Importance of Canonical Product Attributes on User Opinions: an Empirical Investigation

Tiago de Melo
Universidade do Estado do Amazonas
Manaus, Brazil
tmelo@uea.edu.br

Altigran S. da Silva
Universidade Federal do Amazonas
Manaus, Brazil
alti@icomp.ufam.edu.br

Edleno S. de Moura
Universidade Federal do Amazonas
Manaus, Brazil
edleno@icomp.ufam.edu.br

Pável Calado
Universidade de Lisboa
Lisboa, Portugal
pavel.calado@tecnico.ulisboa.pt

ABSTRACT

When making purchasing decisions, customers usually rely on information from two types of sources: product specifications, provided by manufacturers, and reviews, posted by other customers. Both kinds of information are often available on e-commerce websites. While researchers have demonstrated the importance of product specifications and reviews as separate and valuable sources to support purchase decision-making, a largely uninvestigated issue is what is the relationship between these two kinds of information. In this paper we present an empirical study on the use of direct and indirect mentions to canonical product attributes, that is, those defined by manufactures in product specifications, in the reviews written by customers. For this study, we analyzed more than 1,100,000 opinionated sentences available in about 650,000 user reviews from Amazon.com across five product categories. Our results indicate that user opinions are indeed guided by the attributes from product specifications and highlight the influence of canonical attributes on the user reviews.

CCS CONCEPTS

• **Information systems** → **Sentiment Analysis**; *Web mining*.

KEYWORDS

User Reviews, Product Attributes, Opinion Mining

ACM Reference Format:
Tiago de Melo, Altigran S. da Silva, Edleno S. de Moura, and Pável Calado. 2019. The Importance of Canonical Product Attributes on User Opinions: an Empirical Investigation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3316458

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3316458

1 INTRODUCTION

In typical e-commerce Web sites, descriptions of products in the catalog usually consist of objective (factual) data provided by manufacturers informing customers of product's characteristics, which are represented as a set of previously defined product attributes. For instance, for laptops, the brand, the weight and the processor model are commonly available to help potential customers make their purchase decisions. These attributes are called here *canonical*. However, with the rise of the so-called Web 2.0, there is also a large amount of subjective (opinionated) information available on products and their characteristics. In most cases, this subjective information is provided by opinions issued by other customers in reviews.

The importance of considering subjective information in addition to objective (factual) information has been verified in many e-commerce related applications. Indeed, considering opinions issued by other people before purchasing a product is a common practice, especially since there are plenty of opinions available on the Web. In a recent survey [15], it was found that 82% of the Americans refer to online reviews when they first purchase a product, and 40% always refer to online reviews when purchasing good. Another comprehensive survey of online shoppers from 5 different continents [10] revealed that 45% of consumers consider reviews as the most influential aspect of social media for their online shopping behavior. This survey also shows that checking reviews about products or retailers is the fourth most common activity of in-store shoppers with mobiles/smartphones.

On the other hand, product attributes also comprise valuable sources to support purchase decision-making. According to Park et al. [8], product attributes in websites encourage consumer browsing behavior, which can often lead to impulse buying behavior. Therefore, product attributes are a crucial element that influences customer product choice [2].

Despite the substantive importance of user opinions and product attributes for customer decisions, the relationship between these two kinds of information has been largely overlooked. Motivated by the above observations, in this paper, we empirically study the importance of product attributes on user opinions. Specifically, we analyze the use of direct and indirect mentions to canonical product attributes, that is, those defined by manufactures in product specifications, in the reviews written by customers. Our ultimate goal is

to extend previous research studies that assessed the importance of product attributes and user opinions in a separate way. To this end, we executed an extensive experimental evaluation using about 650,000 real reviews, composed of more than 1,100,000 opinionated sentences, and more than 30,000 products in five different categories of electronic products (cameras, cell phones, DVDs, laptops, and routers). The results from this study revealed the importance of canonical product attributes on the user opinions.

The remainder of this paper is organized as follows. Section 2 provides a brief review of related work. In Section 3, we describe hypotheses we have formulated for this study. Then, we describe the experimental dataset we used to verify our hypotheses in Section 4 and the main results and findings that support our hypotheses in Section 5. Finally, we conclude by discussing our results and by proposing directions for future research in Section 6.

2 RELATED WORK

In this section, we describe related work been done on the importance of product attributes [3, 6, 13, 17] and user reviews on purchasing decisions [1, 2, 12, 14, 16].

There are some studies which investigated the importance of product attributes in e-commerce domain. Lee and Nguyen [3] investigated the importance of product attributes in purchasing fashion goods and the influence of these attributes on preferences for external versus local fashion brands. Rahul Rai [13] proposed a method that identifies key product attributes from online customer reviews and ranks each product's attributes with part-of-speech (POS) tagging. Maslowska et al. [6] studied how review characteristics (i.e., number of reviews), product characteristics (i.e., price) and customer behaviors (i.e., reading reviews) interact with each other to influence purchase decisions. Wang et al. [17] conducted a study on online reviews to measure how product attributes impact customer satisfaction.

On the other hand, several works have investigated the importance of customer reviews. Kostyra et al. [2] investigated the impact of online customer reviews on customer's decisions. Qazi et al. [12] investigated why some reviews are more helpful compared to others. Jo and Oh [1] proposed two distinct models to discovering what aspect expressions are evaluated in user reviews and how sentiments for different aspects are expressed. Sun et al. [16] proposed a method of using external user generated data to evaluate the relative importance of an entity's attribute. Singh et al. [14] developed a method based on machine learning that can predict the helpfulness of the consumer reviews using several textual features such as polarity, subjectivity, entropy, and reading ease.

While prior studies have contributed to understanding on the importance of product attributes and user reviews as separate and valuable sources to support purchase decision-making, our work investigated the relationship between two kinds of information.

3 RESEARCH HYPOTHESES

In this study, we aim at investigating the importance of canonical product attributes on user opinions. We choose online user reviews on electronic products available in e-commerce websites as a context for our research due to the popularity and importance of this domain.

A review is a text posted by a user on an e-commerce website, usually reporting their experience with a specific product, which we call the *target product* of the review. Each review is composed of a set of *sentences*. Sentences that express factual information are called *objective* sentences, while sentences that express personal feelings or beliefs are called *subjective* or *opinionated* sentences [5]. We are interested in the latter because they represent the user's opinions on a product.

An opinionated sentence can be further classified as *comparative* or *direct*. A comparative sentence expresses an opinion on similarities or differences between two or more products. The sentence *"the camera of the iPhone is much better than Galaxy"* is an example of a comparative sentence. A direct opinionated sentence (DOS) expresses an opinion directly on a characteristic or part of the product, or on the product as a whole. The sentence *"the camera of the iPhone is fantastic"* is an example of a direct opinion. As our goal is to investigate the importance of canonical product attributes regarding the *target product*, we only consider direct opinionated sentences, since opinions in comparative sentences are relative.

For online user reviews posted on e-commerce websites, we expect to obtain a huge volume of DOS. We thus formulate the following hypotheses:

$\mathcal{H}1$**.** E-commerce websites are a valuable source of opinions on *target products*.

When analyzing real user reviews, we noticed that reviews also include opinions that do not refer to a canonical product attribute, but to the product as a whole. Furthermore, opinions may also refer to attributes that are not represented in the product catalog. Thus, we consider that opinions may have three distinct targets: a) `Attribute`, when opinion is on a canonical product attribute; b) `General`, when opinion is on the product as whole; and c) `Other`, when opinion is on a characteristic of the target product that is not represented as a canonical attribute. In spite of that, we do expect that canonical attributes have a persuasive effect on online user reviews. Hence, we hypothesize:

$\mathcal{H}2$**.** Most of the user opinions posted in e-commerce web sites is on canonical product attributes.

$\mathcal{H}3$**.** According to the user opinions, there are certain canonical attributes that are more relevant than the other attributes.

$\mathcal{H}4$**.** Customers often make either direct and indirect mentions to canonical product attributes.

To verify our hypotheses, we built an experimental dataset that will be explained in the next section.

4 EXPERIMENTAL DATASET

We built our experimental dataset from a large collection of about 142 million reviews previously crawled from the Amazon.com web site [7][1], and selected all reviews from each of five categories: cameras (CAM), cell phones (CEL), dvd players (DVD), laptops (LAP), and internet routers (ROT). Each review in this collection identifies the product to which it refers. Table 1 presents, for each category, the number of reviews and sentences in this collection, along with the number of products referred in the reviews.

[1] Available at http://jmcauley.ucsd.edu/data/amazon

Table 1: Summary of Amazon.com review collection.

Category	No. products	No. reviews	No. sentences
CAM	8,893	203,836	1,012,077
CEL	7,693	182,491	707,407
DVD	2,503	61,836	243,939
LAP	9,491	115,138	580,955
ROT	1,592	84,059	329,305
Total	**30,172**	**647,360**	**2,864,683**

To complete our experimental dataset, we crawled from Amazon.com the set of canonical attributes used in products of each of the five categories. These sets are shown in Table 2.

Table 2: Canonical product attributes.

Category	Product Attributes
CAM	*Dimension*, *Exposure Control*, *Imaging Memory*, *Performance*, *Power*, *Price*, *Zoom*
CEL	*Battery*, *Camera*, *Dimension*, *Display Memory*, *Price*, *Processor*, *Software*
DVD	*Accessory*, *Audio*, *Dimension*, *Price Sound*, *Video*
LAP	*Battery*, *Connectivity*, *Dimension*, *Graphic Memory*, *Price*, *Processor*, *Screen*, *Software*
ROT	*Accessory*, *Coverage Area*, *Dimension Ports*, *Price*, *Security*, *Software*, *Speed*

Recall that our study targets only direct opinionated sentences (DOSs). Thus, we proposed a method named *filterDOS* to select DOSs from reviews.

filterDOS is carried out in three steps and it allows to identify three types of sentences: *factual*, *comparative*, and *DOS*. In the first step, the reviews are broken down into sentences. In the second step, the method identifies subjective sentences obtained in the first step. Objective (factual) sentences are discarded. As an example, the sentence "*I had to go back to store and they did not have the device stocked*" written by a user is not opinionated and therefore it should be discarded. For this task, we implemented the unsupervised method proposed by Qadir [11], where typed dependency relations, such as open clausal complements or adjectival complements, are used for identifying subjective sentences. The subjective sentences can be further classified into *comparative* or *direct* sentence. A comparative sentence expresses a relation of similarities or differences between two or more products. As an example, the sentence "*the camera of the iPhone is much better than Galaxy*" is a comparative sentence. A direct opinionated sentence (DOS) expresses an opinion directly about a characteristic or part of the product, or the product as whole. The sentence "*the camera of the iPhone is fantastic* is an example of DOS. Finally, in the last step, *filterDOS* eliminates the comparative sentences, based on the unsupervised method proposed by Liu [4].

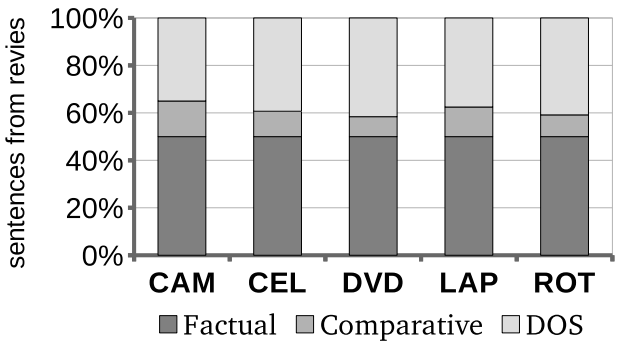


Figure 1: Percentage (%) of sentences of each type in user reviews.

Figure 1 presents the results obtained using *filterDOS* on the Amazon.com review collection. We can observe that, in average, 51.31% of the sentences are factual, 40.05% are DOSs, and only 8.64% are comparative. It is noticeable that there are very few of comparative sentences in product reviews. This is due to the fact that e-commerce site users focus on writing only about the product of interest, unlike what occurs, for example, in forums where users usually write comments comparing products.

The initial review collection described in Table 2 has more than 2,800,000 sentences. Therefore, even after filtering factual and comparative sentences, the number of sentences to be analyzed is still huge, with more than a one million sentences.

A central concept in opinion mining that we leverage in our study is that of *aspect expressions*. We consider that an opinion is represented by an aspect expression [5], where an aspect is any reference made in an opinion to a particular part or characteristic of the product, or even to the product as a whole. Since it would be unfeasible to manually annotate all aspects found this huge volume of sentences, we created an experimental dataset focused on the 100 most frequent aspect expressions found in the DOS of each product category. We argue that in a practical setting, handling a few top frequent aspect expressions is more valuable than considering every single aspect expression from a potentially huge list.

To select the 100 most frequent aspect expressions, we first run the aspect extraction method proposed by Poria et al. [9]. This was implemented and all possible aspect expressions identified by this method were extracted. We then ranked these expressions according to their frequency. To assure that we only use true aspect expressions, we manually inspected the extracted expressions using the ranking order, and removed those that we did not consider as aspect expressions. In the end, only the 100 most frequent true aspect expressions were kept for each product category. We named this set of 100 aspect expressions as *top 100 aspects*.

Finally, for each product category, we manually examined each of the *top 100 aspects* selected previously, and annotate each one with the canonical product attributes that is most related to it. We also annotate accordingly cases where the opinion is on the product

as whole (General) and when opinion is on a characteristic of the target product that is not represented as canonical attribute (Other). This annotated dataset was then used to verify our formulated hypotheses ($\mathcal{H}1$, $\mathcal{H}2$, $\mathcal{H}3$, and $\mathcal{H}4$). Table 3 presents a summary of the experimental dataset we generated, which shows, for each category, the total number of DOSs (No. DOSs) and the total number of DOSs that include at least one of the top 100 aspects (No. DOSs *top100*). As it can be observed, these DOSs represent more the 50% of all DOSs.

Table 3: Summary of our experimental dataset.

Category	No. DOSs	No. DOSs *top100*
CAM	476,605	249,714
CEL	277,712	138,939
DVD	89,525	48,608
LAP	189,782	126,865
ROT	123,336	73,027
Total	**1,156,960**	**637,153**

5 RESULTS

In this section, we present the main results and findings that support our hypotheses formulated on the importance of canonical product attributes on user opinions.

5.1 Distribution of Sentences among Kinds of Targets

To test the hypothesis $\mathcal{H}1$, we created the method *filterDOS* to select direct opinionated sentences from reviews. According to our experimental dataset presented in Table 3, more than 1,100,00 sentences are DOSs. The fact that a large fraction of the sentences is DOS supports our hypothesis $\mathcal{H}1$ that e-commerce websites, such as Amazon.com, are indeed useful as a valuable source of opinions on target products. Furthermore, from the experimental dataset, we can observe that more than 55% DOSs contain at least one of the *top 100 aspects*. This finding corroborates our assumption that handling a few top frequent aspect expression is more valuable than showing every single aspect expression from a potentially huge list.

Figure 2 summarizes the distribution of sentences among the three kinds of targets: Attributes, General, and Other. The number of sentences containing at least one of the *top 100 aspects* that form the opinions are annotated with a kind of target. As explained in Section 3, a single sentence may contain more than one opinion, and each opinion can refer to a different kind of target. Thus, the sum of percentage of all kinds of targets may be greater than 100%.

Again we observe that most of the DOSs include aspect expressions that refer to canonical product attributes, identified as Attribute. For example, in CAM and LAP categories they account for more than half of the sentences. Also, a large share of the sentences contain opinions referring to the target General.

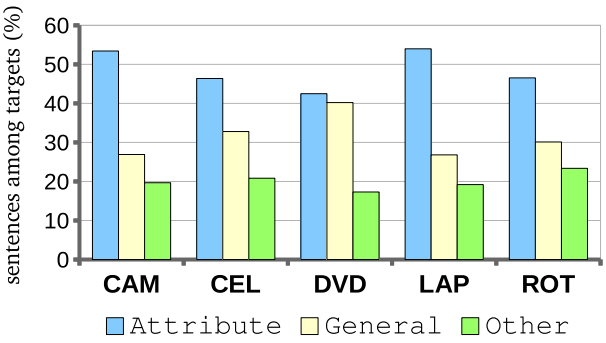


Figure 2: Distribution of sentences among targets.

5.2 Distribution of Aspect Expressions

Figure 3 shows the distribution of the *top 100 aspects* among the three kinds of targets. Notice that for all categories the fraction of aspect expressions that represent the target Attribute is higher than 50%. This supports our hypothesis $\mathcal{H}2$, that most of the user opinions posted in e-commerce web sites is about the canonical product attributes. In addition, it can be observed that a larger share of the aspect expressions refer to the target Other. For example, in CAM, CEL and DVD, almost one quarter of the aspect expressions are in opinions which were annotated as the target Other.

An intriguing problem we left for future work is to further analyze cases such as these to look for specific latent characteristics that, although are not represented by some canonical attribute, are of interest for users. For example, "keyboard" is the second most frequent aspect expression in LAP, but typically, there is no attribute referring to it in the canonical product attributes.

In sum, Figure 3 suggests that users comment more frequently on the specific characteristics of the products than on the product as a whole. This shows the relevance of properly addressing references to attributes in user reviews.

5.3 Distribution of Sentences among Canonical Product Attributes

Figure 4 shows the distribution of sentences among the canonical product attributes for each category. In these graphs, each vertex in the polygon represents a canonical product attribute defined by manufacturers in product specifications. The graph shows the percentage of sentences that contain an aspect expression that corresponds to a given attribute. In each graph, the canonical attributes are placed in clockwise order, from the most to the least frequently referred. For example, more than 40% of the sentences that include at least one of the *top 100 aspects* in the DVD category refer to the attribute *Accessory*.

There are some attributes that are much more frequently referred to in reviews than other from the same category. For example, in the CEL category, users comment 2 times more on Battery than on the Price of cell phones. These experimental results support our hypothesis $\mathcal{H}3$ that there are certain canonical attributes that are more relevant than the other. Interestingly, in the five categories in this experiment, the price is not the most commented attribute.

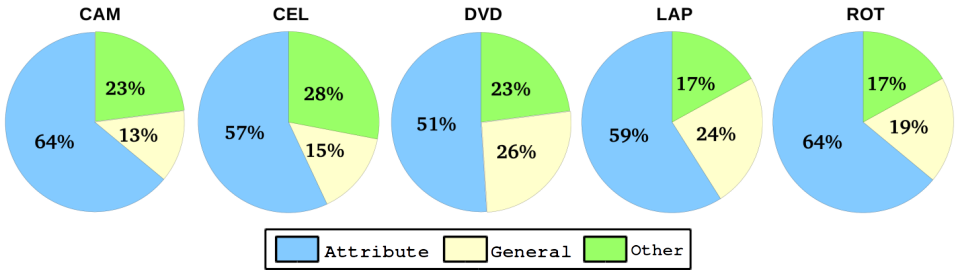
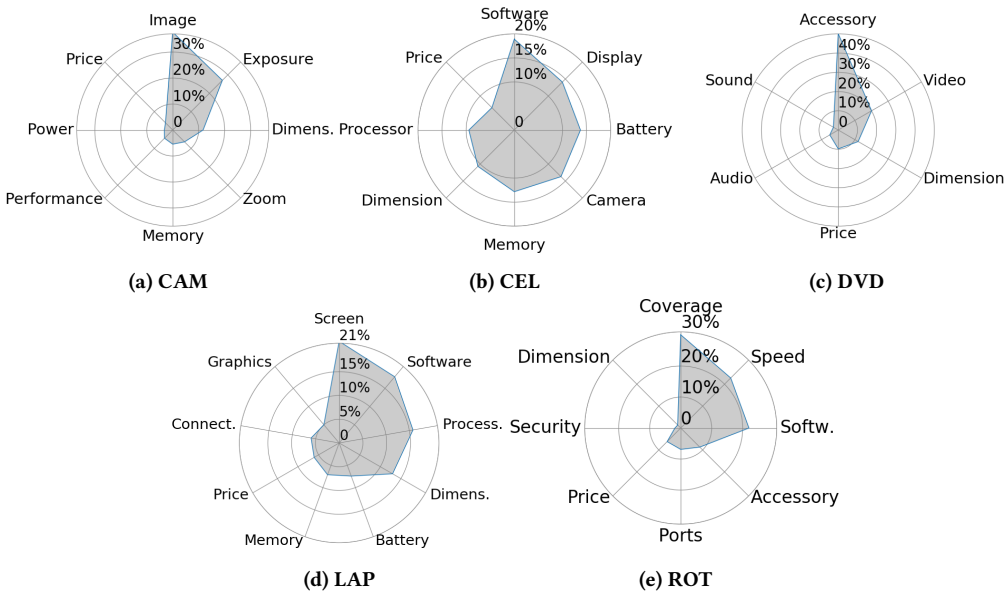


Figure 3: Distribution of the top 100 aspects among the three kinds of targets.



(a) CAM **(b) CEL** **(c) DVD**

(d) LAP **(e) ROT**

Figure 4: Distribution of sentences among the attributes they refer to. Labels are canonical product attributes.

5.4 Diversity of Aspect Expressions over Attributes

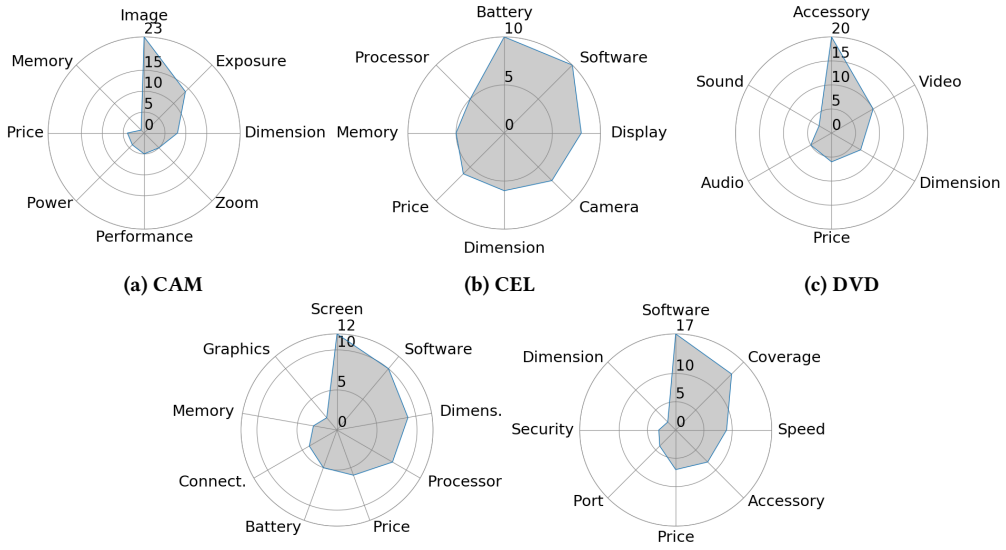
Figure 5 shows the distribution of aspect expressions extracted from user reviews over canonical product attributes in each category. In these graphs, we show the quantity of unique aspect expressions that refer to the same attribute. For example, in the LAP category, we found ten different aspect expressions that refer to the attribute *Software*. Analyzing the sentences, we found that users do indeed employ several different terms such as "apps", "system", "vista", and "program" to refer the attribute *Software* in the LAP category. This experimental evaluation supports our hypothesis $\mathcal{H}4$ that customers often make either direct and indirect mentions to canonical product attributes.

To give an idea of the *top 100 aspects* mentioned, Table 4 illustrates the ten most frequent aspect expressions extracted in the reviews of each category, along with the attribute name, when they refer to `Attribute` target, or the target name (`General` or `Other`). From these results is it apparent that the ten most frequent aspect expressions extracted are quite representative of each product category and, more importantly, the results show which are the most commented aspects related to canonical attributes. Notice that, most aspect expression do not match exactly with the name of the product attribute.

6 CONCLUSIONS

The main goal of this research is to investigate the importance of canonical product attributes on user opinions. Based on a empirical evaluation carried out over a representative collection of real user reviews, we were able to verify hypotheses we have formulated on this issue. Our results were drawn from a large experimental dataset we built with more than 1 million direct opinionated sentences (DOSs) in five product categories.

In our study, we verify that a large fraction of sentences in reviews is composed of direct opinionated sentences and this validate our hypothesis that e-commerce websites are a valuable source of opinions on target products ($\mathcal{H}1$). We use this large number of sentences and, by means of a well defined protocol, we could verify that the most of user opinions posted in e-commerce web sites is on one of the canonical product attributes ($\mathcal{H}2$). Furthermore, we could verify that there are certain canonical attributes that are



(a) CAM **(b) CEL** **(c) DVD**

(d) LAP **(e) ROT**

Figure 5: Distribution of different aspect expressions according the canonical product attributes presented in Table 2.

Table 4: The 10 most frequent aspect expressions in reviews of each category from the Amazon dataset.

CAM	CEL	DVD	LAP	ROT
camera (General)	phone (General)	unit (General)	laptop (General)	router (General)
quality (General)	easy (Other)	dvd (General)	keyboard (Other)	easy (Other)
picture *(Imaging)*	quality (General)	easy (Other)	computer (General)	instructions (Other)
lens *(Exposure Control)*	feature (Other)	quality (General)	software *(Software)*	software *(Software)*
shots *(Exposure Control)*	card *(Memory)*	player (General)	fast *(Processor)*	unit (General)
features (Other)	size *(Dimension)*	picture quality *(Video)*	size *(Dimension)*	speed *(Speed)*
size *(Dimension)*	software *(Software)*	instructions (Other)	card *(Memory)*	device (General)
card *(Memory)*	camera *(Camera)*	features (Other)	screen *(Screen)*	internet *(Coverage Area)*
settings (General)	screen *(Display)*	picture *(Video)*	easy (Other)	network *(Coverage Area)*
pics *(Imaging)*	keyboard (Other)	product (General)	graphics *(Graphics)*	settings (Other)

more relevant for users than the other attributes ($\mathcal{H}3$). Finally, we could conclude that customers often make either direct and indirect mentions to canonical product attributes using several distinct expressions ($\mathcal{H}4$).

This study contributes to understanding the importance of canonical product attributes on user opinions and our results indicate that user opinions are indeed guided by the attributes from product specifications and highlight the influence of canonical attributes on the user reviews.

Some limitations are associated with this study, which, however, can provide directions for future research. First, we considered only aspect expressions to represent user opinions. Future research could extend the current study for examining other components of opinions, such as star ratings of reviews, opinion polarity, and opinion posting time. Second, our analysis is restricted to products, which in turn, have well established set of canonical attributes provided by manufacturers. However, domains such as restaurants or hotels do not have clear canonical attributes. Therefore, a future research could extend the current study to these domains.

REFERENCES

[1] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 815–824.

[2] Daniel S Kostyra, Jochen Reiner, Martin Natter, and Daniel Klapper. 2016. Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing* 33, 1 (2016), 11–26.

[3] Jaeha Lee and Minhthu Jill Nguyen. 2017. Product attributes and preference for foreign brands among Vietnamese consumers. *Journal of Retailing and Consumer Services* 35 (2017), 76–83.

[4] Bing Liu. 2010. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing* 2 (2010), 627–666.

[5] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

[6] Ewa Maslowska, Edward C Malthouse, and Vijay Viswanathan. 2017. Do customer reviews drive purchase decisions? The moderating roles of review exposure and price. *Decision Support Systems* 98 (2017), 1–9.

[7] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[8] Eun Joo Park, Eun Young Kim, Venessa Martin Funches, and William Foxx. 2012. Apparel product attributes, web browsing, and e-impulse buying on shopping websites. *Journal of Business Research* 65, 11 (2012), 1583–1589.

[9] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A Rule-Based Approach to Aspect Extraction from Product Reviews. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*. 28.

[10] PwC. 2016. They say they want a revolution – Total Retail 2016. http://pwc.to/2kRAxYL. Accessed 08 September 2018.

[11] Ashequl Qadir. 2009. Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Proceedings of the Workshop on Events in Emerging Text Types*. Association for Computational Linguistics, 38–43.

[12] Aika Qazi, Karim Bux Shah Syed, Ram Gopal Raj, Erik Cambria, Muhammad Tahir, and Daniyal Alghazzawi. 2016. A concept-level approach to the analysis of online review helpfulness. *Computers in Human Behavior* 58 (2016), 75–81.

[13] Rahul Rai. 2012. Identifying key product attributes and their importance levels from online customer reviews. In *ASME 2012 international design engineering technical conferences and computers and information in engineering conference*. American Society of Mechanical Engineers, 533–540.

[14] Jyoti Prakash Singh, Seda Irani, Nripendra P Rana, Yogesh K Dwivedi, Sunil Saumya, and Pradeep Kumar Roy. 2017. Predicting the âĂIJhelpfulnessâĂİ of online consumer reviews. *Journal of Business Research* 70 (2017), 346–355.

[15] Aaron Smith and Monica Anderson. 2016. Online shopping and e-commerce. https://www.pwc.es/es/publicaciones/retail-y-consumo/assets/total-retail-2016.pdf. Accessed 10 September 2018.

[16] Shengjie Sun, Dong Yang, Hongchun Zhang, Yanxu Chen, Chao Wei, Xiaonan Meng, and Yi Hu. 2018. Important Attribute Identification in Knowledge Graph. *arXiv preprint arXiv:1810.05320* (2018).

[17] Yuren Wang, Xin Lu, and Yuejin Tan. 2018. Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electronic Commerce Research and Applications* 29 (2018), 1–11.

Car Theft Reports: a Temporal Analysis from a Social Media Perspective

Juglar Diaz
University of Chile
Santiago, Chile
juglar.diaz@ug.uchile.cl

Barbara Poblete
University of Chile
Santiago, Chile
barbara@poblete.cl

ABSTRACT

Complex human behaviors related to crime require multiple sources of information to understand them. Social Media is a place where people share opinions and news. This allows events in the physical world like crimes to be reflected on Social Media. In this paper we study crimes from the perspective of Social Media, specifically car theft and Twitter. We use data of car theft reports from Twitter and car insurance companies in Chile to perform a temporal analysis. We found that there is an increasing correlation in recent years between the number of car theft reports in Twitter and data collected from insurance companies. We performed yearly, monthly, daily and hourly analyses. Though Twitter is an unstructured source and very noisy, it allows you to estimate the volume of thefts that are reported by the insurers. We experimented with a Moving Average to predict the tendency in the number of car theft reported to insurances using Twitter data and found that one month is the best time window for prediction.

CCS CONCEPTS

• **Information systems** → **Social networks**; *Spatial-temporal systems.*

KEYWORDS

Twitter, Temporal patterns, Car thefts

ACM Reference Format:
Juglar Diaz and Barbara Poblete. 2019. Car Theft Reports: a Temporal Analysis from a Social Media Perspective. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.3316462

1 INTRODUCTION

Social media has been the source of information in multiple studies for understanding events in the physical world. The study of complex human activities such as criminal acts, in particular car theft, requires the analysis of multiple sources of information. In this context, social networks can contribute with valuable information that can not be found in other sources. Twitter[1] is one of the most popular social networks in Chile. In Twitter opinions and statements are shared in real time through messages with a maximum of 140 characters, called tweets. People share different types of information in Twitter including reports of car theft with the hope that it will help them to recover the stolen car. Tweets have an associated timestamp corresponding to the moment where the tweet was created. This means that temporal patterns of car theft can be extracted from information published on Twitter. Also, car insurance companies collect data about car theft reported by their clients. The objective of this work is to study temporal patterns of car theft from two perspectives: social media car theft reports, specifically Twitter and car theft insurance records (CIR).

Many have studied how to use Twitter to better understand events in the physical word [5, 6, 9, 11, 12, 16, 17]. Temporal patterns have been analyzed to study activities on urban environments [16], mobility models [17], event detection [12] and forecasting [5]. Most of the works in the intersection of crime incidents and social media data have focused on enriching event prediction models based on historical records with information extracted from Twitter. Patterns that have been extracted as Twitter indicators are Sentiment Analysis [10] and Topic Modeling [1].

In this work we use Twitter to better understand the car theft phenomenon with an emphasis on temporal patterns. With that end we study yearly, monthly, weekly, daily and hourly patterns using the two datasets. Twitter car theft reports are a sample of the whole population. Also, Twitter reports covers both insured and not insured cars. As CIR only covers those cars that are insured, the two datasets are complementary. Twitter data is freely available and produced in real time which allows analysis and predictions in the moment.

Resuming in this work we explore temporal patterns of car theft reports in Twitter and insurances. We explore frequency patterns for recent years in both datasets and study the number of daily and hourly reports. Finally we explore how we can predict the tendency of number of reports in the insurances from Twitter data. Our findings in this paper are:

- In recent years there have been an increasing correlation between the number of car theft reports in Twitter and the number of car theft reports in insurance companies.
- Twitter, despite being an unstructured source and very noisy, allows us to estimate the volume of theft that are reported in the insurance companies.
- Most of the robberies take place late in the night with a difference between weekends and weekdays. For weekdays robberies are more common in the morning while for weekends is more common the early morning.

[1] https://twitter.com/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3316462

- The best time window estimator for the number of car theft reports for the insurers using Twitter data is a one-month time window.

It is worth mentioning that this work is complementary to that presented by [13] (In Spanish) in the sense that we use the same datasets as them but our analysis is entirely new. The rest of this paper is organized as: Section 2 describes related works aiming to better understand crime incidents. Section 3 presents the description of the experiments and finally in Section 4 we present the conclusions.

2 RELATED WORKS

In this section we describe works directly related to temporal patterns of car theft. Then we describe works that use social network data to alleviate and better understand crime incident related problems.

Car theft: In [4] Chen et al. analyzed car theft report data from Taiwan and discovered that the sooner the robbery is reported, the more likely it is that the car is recovered. In [3] Chen focused their analysis on the time elapsed after the incident was notified. They found that criminals prefer to commit crimes between 4:00 AM and 8:00 AM.

Social networks and Crime incidents: The main focus to enhance crime incident applications with social network data has been enriching predictive models based on historical records with information extracted from Twitter. Patterns extracted as Twitter indicators are Sentiment Analysis [10] and Topic Modeling [1]. In the work of [2] the authors use Twitter data from the ten most violent cities and the ten least violent cities in the United States. In their findings less violent cities presented a lower proportion of tweets associated with violent acts than the rest of the cities. Also, cities with the highest rates of violence presented a higher intensity of negative tweets, although some of the less violent ones also presented high levels of tweets with negative sentiments. In [5] Chen et al. proposed a crime prediction model that indicates when and where a crime will occur. Their experiments showed that the inclusion of sentiment information from Twitter improved the quality of the model. The work reported in [7] aims to create a model to predict the probability that a certain type of crime occurs in a certain place for the next day. Gerber compared two types of models, the first uses only historical records of crimes, the second includes the data from Twitter, which is processed using topic modeling. Of the 25 types of crimes considered, 19 showed improvements in the prediction when the topics of Twitter were added to the predictive model. The goal of Wang et al. in [15] is to predict whether a crime incident of hit-and-run will happen on day d considering the topics discussed on Twitter on day $d-1$. To add the information from Twitter to the prediction model [15] follows two steps: first the tweets are processed by a Semantic Role Labeling [8] tool and then topics are extracted using LDA [1]. The model performed better than a baseline based on a uniform distribution. In [13] Vásquez et al. present an analysis similar to our proposal in the sense that they also compare car theft reported in Twitter with car theft records from insurances. Their main findings are that robberies reported on Twitter present a higher rate of recoveries and that in general cars reported on Twitter are of lower value than those reported by insurance companies.

Our work is different from the ones we described in this section in that we analyze temporal patterns of different granularities. We study yearly, monthly, daily and hourly patterns in two related datasets: car theft reports from insurance companies and Twitter car theft reports both in Santiago, Chile and in the same period of time.

3 EXPERIMENTS

Our goal is to discover new information and similarities about car theft from two different data sources. In particular, we focus on what temporal patterns we can discover from car theft reports on the social network Twitter in comparison to data of car theft reports from insurance companies in Chile.

3.1 Data

For our experiment we used two datasets of car theft reports in Chile from Jan 1, 2012 to Jan 31, 2016. One was collected from Twitter and the other from insurance companies. The Twitter dataset was collected querying to *https://twitter.com/search-home*. The query was: *[(robar | robado | robaron | robo | robada) & patente & near:Chile]* it translates to *[(steal | stolen | stolen | steal | stolen) & license plate & near:Chile]* . After deleting duplicates and data cleaning resulted in a dataset of around 7,000 records. The dataset from insurance companies consists of around 40,000 car theft reports submitted by users to car insurance companies in Chile. The timestamp is an estimation of when the incident occurred according to the client. The dataset was provided to us by the National Association of Car Insurance Companies through a collaboration agreement.[2]

3.2 Analysis of Temporal Patterns

Human activities are mostly associated to temporal patterns with different granularities as day of the week and time of day. In the morning children are taken to school and the movement of people occurs from home to the workplace. Family activities usually take place on weekends, while afternoons and evening activities are more related to leisure and nightlife in the city. These types of activities can be associated with patterns of car theft.

First we show in Figure 1 the number of robberies per month normalized according to the total (for each dataset) from 2012 to 2016. We can see that there is a great correspondence between the number of reports from Twitter and from CIR. There is a correlation of 0.42.

Later we disaggregate the graph from Figure 1 by years: 2013, 2014, 2015 and 2016 in Figure 2. By separating the data for years we can observe an increase in the correlation between the volume of thefts reported on Twitter and CIR in recent years. All this indicates that Twitter, despite being an unstructured source and very noisy, allows us to estimate the volume of car theft that are reported by the insurers.

A similar relation is found regarding the distribution of thefts by days of the week. Figure 3 shows the number of thefts per day of the week of all the reports in both data sources. In general,

[2]Project FONDEF project ID16I10222, CONICYT.

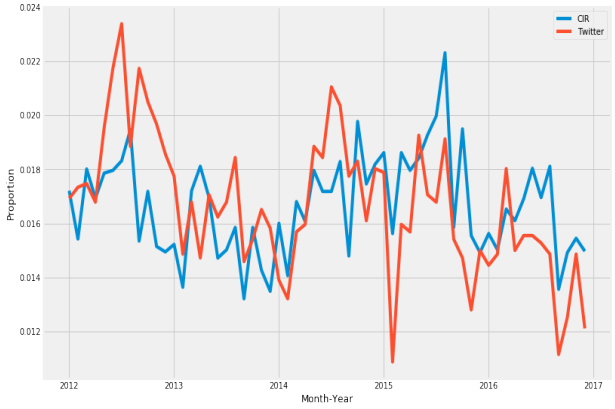


Figure 1: Number of car theft per month normalized according to the total from 2012 to 2016. Correlation 0.42.

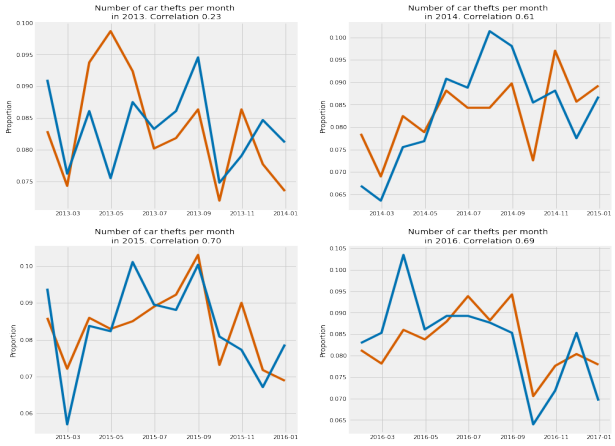


Figure 2: Number of car theft per month normalized according to the total for years 2013, 2014, 2015 and 2016.

thefts increase until mid-week, gradually decreasing towards the weekend.

Figures 4 and 5 show the normalized number of robberies by day of the week and hour of the day. The information associated with the time in both data sources is not the real moment when the robbery occurred. In the case of Twitter, it is when the tweet is created, while in the case of CIR it is the time of the robbery reported by the owner of the car and it is expected to be an approximated value.

In CIR most cases are concentrated late at night and early in the morning, while on Twitter they concentrate in the afternoon. This behavior is consistent with the fact that in CIR the timestamp is the timestamp reported by the owner, while in Twitter it is the time when the tweet is written. This may be due to the victims having more time to report the theft on Twitter during these hours. In the case of CIR on weekends there is an increase at dawn, while that on weekdays this increase is transferred to the morning hours.

From these exploratory analysis we can see that Twitter, despite being an unstructured source and very noisy, allows for an estimate

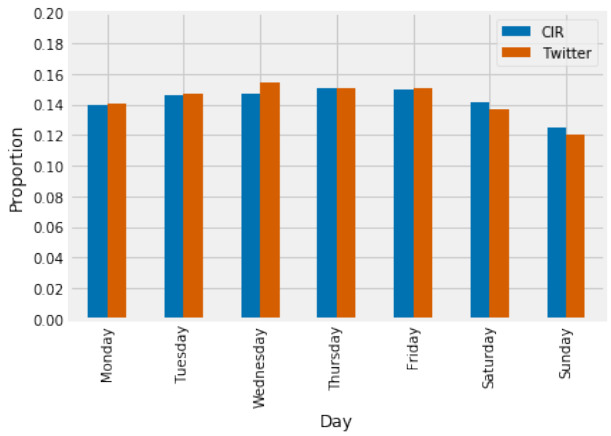


Figure 3: Number of car theft reports by day of the week.

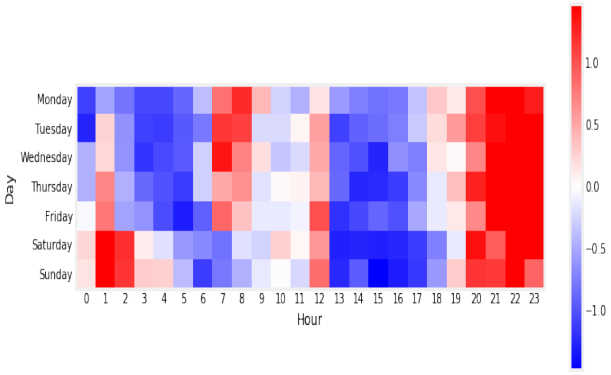


Figure 4: Normalized number of car theft reports by day of the week and hour of the day in CIR.

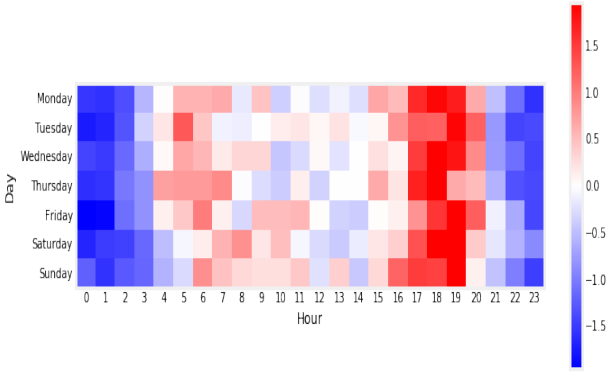


Figure 5: Normalized number of car theft reports by day of the week and hour of the day in Twitter.

of the volume of thefts that are reported to the insurers. Finally we study the best time window to make predictions about the tendency of car theft reports in the CIR using Twitter data. For that

purpose we use a Moving Average (MA) from Twitter to predict the tendency of the car theft reports in CIR. We use the Pearson correlation coefficient as evaluation metric. In Figure 6 we show MA for 1,2,3,4 and 5 months. We can see that the best correlation between the MA and CIR is using the last month as predictor. In Figure 7 we show a finer granularity of the prediction by predicting the number of reports weekly. We can see that despite testing with fewer (1,2,3) weeks than a month, still the mean of one month of reports in Twitter is the best predictor of the tendency for the number of car theft reports in CIR.

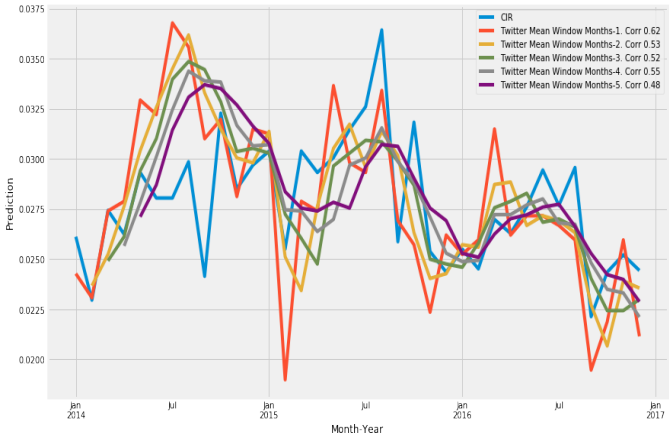


Figure 6: Monthly Moving Average from Twitter to predict the tendency of the car theft reports in CIR.

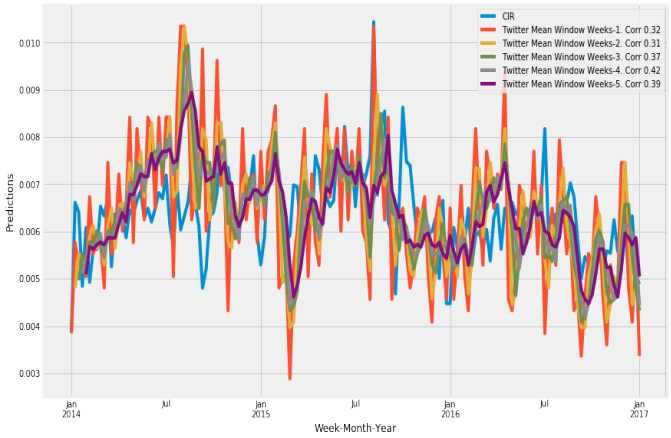


Figure 7: Weekly Moving Average from Twitter to predict the tendency of the car theft reports in CIR.

4 CONCLUSIONS

In this work we studied the car theft phenomenon from a social media perspective. We studied temporal patterns and found that there is an increasing correlation in recent years between the number of car theft reports in Twitter and data collected from insurance companies. Twitter, despite being an unstructured source and very noisy, allows us to estimate of the volume of thefts that are reported to the insurers. We experimented with a Moving Average to predict tendency in the number of reports using Twitter data and found that one month is the best predictor.

As future work we are interested in predicting not only the tendency of the number of reports, but the exact number of reports. Also, we would like to enrich the analysis with general information from Twitter and study finer granularities of prediction like predicting the number of thefts by branches.

ACKNOWLEDGMENTS

The work of Juglar Diaz was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160142.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[2] Raja Ashok Bolla. 2014. Crime pattern detection using online social media. (2014).
[3] Patrick S Chen. 2008. Discovering Investigation Clues through Mining Criminal Databases. In *Intelligence and Security Informatics*. Springer, 173–198.
[4] Patrick S Chen, KC Chang, Tai-Ping Hsing, and Shihchieh Chou. 2006. Mining criminal databases to finding investigation cluesâĂŤby example of stolen automobiles database. In *Intelligence and Security Informatics*. Springer, 91–102.
[5] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. 2015. Crime prediction using Twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*. IEEE, 63–68.
[6] Jessica Elan Chung and Eni Mustafaraj. 2011. Can collective sentiment expressed on twitter predict political elections?. In *AAAI*, Vol. 11. 1770–1771.
[7] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.
[8] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics* 28, 3 (2002), 245–288.
[9] Niels Buus Lassen, Rene Madsen, and Ravi Vatrapu. 2014. Predicting iphone sales from iphone tweets. In *Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International*. IEEE, 81–90.
[10] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.
[11] Anshul Mittal and Arpit Goel. 2012. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)* 15 (2012).
[12] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.
[13] Alejandro Vásquez, Rodrigo Joannon, and Richard Weber. 2018. Análisis de redes sociales para mejorar la identificación de patrones de robo de vehıculos. *Revista Ingenieria de Sistemas Volumen XXXII* (2018).
[14] Xiaofeng Wang, Donald E Brown, and Matthew S Gerber. 2012. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*. IEEE, 36–41.
[15] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 231–238.
[16] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 361–370.
[17] Chao Zhang, Keyang Zhang, Quan Yuan, Luming Zhang, Tim Hanratty, and Jiawei Han. 2016. Gmove: Group-level mobility modeling using geo-tagged social media. In *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining*, Vol. 2016. NIH Public Access, 1305.

Characterization of Local Attitudes Toward Immigration Using Social Media

Yerka Freire-Vidal
Social Complexity Research Center
Universidad del Desarrollo
Santiago, Chile
yfreirev@udd.cl

Eduardo Graells-Garrido
Data Science Institute
Universidad del Desarrollo
Santiago, Chile
egraells@udd.cl

ABSTRACT

Migration is a worldwide phenomenon that may generate different reactions in the population. Attitudes vary from those that support multiculturalism and communion between locals and foreigners, to contempt and hatred toward immigrants. Since anti-immigration attitudes are often materialized in acts of violence and discrimination, it is important to identify factors that characterize these attitudes. However, doing so is expensive and impractical, as traditional methods require enormous efforts to collect data. In this paper, we propose to leverage Twitter to characterize local attitudes toward immigration, with a case study on Chile, where immigrant population has drastically increased in recent years. Using semi-supervised topic modeling, we situated 49K users into a spectrum ranging from in-favor to against immigration. We characterized both sides of the spectrum in two aspects: the emotions and lexical categories relevant for each attitude, and the discussion network structure. We found that the discussion is mostly driven by Haitian immigration; that there are temporal trends in tendency and polarity of discussion; and that assortative behavior on the network differs with respect to attitude. These insights may inform policy makers on how people feel with respect to migration, with potential implications on communication of policy and the design of interventions to improve inter-group relations.

ACM Reference Format:
Yerka Freire-Vidal and Eduardo Graells-Garrido. 2019. Characterization of Local Attitudes Toward Immigration Using Social Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316455

1 INTRODUCTION

Migration is a phenomenon faced by many countries, which brings a variety of effects; both in the population from which it emigrates and in the receiving population. One of the effects that worries many countries is intolerance and hostile attitudes toward immigrants. These attitudes have been the focus of many research studies, some of which are focused on individual-level psychological and socio-economic factors [8, 47], and others on the contact between immigrant population and locals [7, 30, 32]. The main methods used in these studies are based on context specific surveys, which makes replication in others societies or countries difficult. The theories that explain the type of attitudes of locals interacting with immigrants can be summarized in two: the Intergroup Contact Theory [1], and the Integrated Threat Theory [40, 49]. The former states that people support multiculturalism and integration. The latter, that people think that immigrants will bring negative effects for their society, including competition for jobs and public services, worsening of the national economy, increase in crime, and the arrival of diseases. Particularly, the attitudes explained by the threat theory can lead to acts of violence, discrimination, and abuse; thus, it is important to understand what factors enhance such attitudes.

However, measuring attitudes is costly and impractical under dynamic scenarios. The most frequent methods are surveys, which are difficult and costly to implement. In this paper, we propose to make use of the information that people publish in Twitter as a proxy of their attitudes toward immigration. It is common to find reactions and attitudes through posts in these platforms, where people express their ideas and opinions voluntarily. We propose to define a spectrum of attitudes based on the two aforementioned theories, and to classify users and tweets into that spectrum. We do so with a semi-supervised topic modeling technique named Topic-Supervised Non-Negative Matrix Factorization [39]. TS-NMF works in a semi-supervised way because some users can be labeled as belonging to each extreme of the spectrum, something that we do with custom-built lexicons for each theory.

We perform a descriptive case study on the Chilean society, because Chile is one of the countries in which migration has reached unprecedented volume in recent years. The statistics show that immigrant population has increased from 0.8% in 1992 to 4.35% in 2017; and where 66.7% of immigrants declare to have arrived mainly in 2016 [31]. For this, Chileans have developed diverse perceptions regarding the number of immigrants in the country and the phenomenon itself. To measure them with our proposed method, we collected more than 206K tweets that discuss immigration in Chile, written by more than 49K users during the year 2017. After inferring user and tweets positions in the spectrum, we performed lexical and network analysis with respect to the spectrum position. In the lexical analysis, we used the "Linguistic Inquiry and Word Count" (LIWC) lexicon [43], typically employed to characterize cognitive and emotional differences in discourse [11, 17, 26]. To analyze the network structure, we estimated the polarization of the retweet and mention networks between users.

As main results, we observed that most of the discussion toward migration in Chile is targeted at Haitian migration, even though other countries have a larger share of the population. We found

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316455

lexical differences in how each attitude discussed migration, and those differences were consistent with theories. For instance, social-related words were correlated with empathetic attitudes, job- and money-related words were correlated with threatening attitudes. In the network, the retweet network was polarized, in coherence as predicted by other studies regarding political discussion [10, 20]. Finally, we notice that the amount and tendency of the tweets (the latter reflects the attitude towards immigration) seems to be influenced by relevant news events on national migration issues. These results can inform public policy designers to improve inter-group relations in the country, as well as increasing the understanding of how people feel regarding an important aspect of globalization.

In summary, the contribution of this paper is two-fold. We proposed a methodology to characterize local attitudes toward migration from tweets. Then, we performed a descriptive case study in Chile using this method, obtaining results that are coherent with social theory, with added depth based on the rich information that can be extracted from Twitter.

This paper is structured as follows. Section 2 discusses the related work. Section 3 describes the social theories that guided our analysis. Section 4 describes the data set we analyzed. Section 5 describes the methodology. Section 6 describes the results of applying the methodology to the data set. Section 7 discusses the implications of our work. Finally, Section 8 states our conclusions.

2 RELATED WORK

Migration is a widely studied topic because there are many issues associated to this phenomenon. Some researchers have focused on studying the economic impacts related to migration [23–25, 47], others on social cohesion [30, 32, 33, 48]. Within these studies, those who have focused on integration [36] and racism/xenophobia stand out [7, 27, 44]. Our work seeks to contribute in the latter area, mainly due to the subject of our case study, Chile, a society that in a short time has faced a massive influx of immigrants. Migration in Chile has been a national issue, causing controversy in presidential elections, news, and municipal institutions. However, measuring attitudes is not a simple problem, nor a solved one. Twitter is currently a platform widely used in studies of human behavior, since it provides a valuable source of data. Studies that have used Twitter have allowed to reveal socio-cultural characteristics of users or societies, including the level of integration of immigrants in a city [36], attitudes in response to triggering events, such as terrorist attacks [13], the influence of culture in personal actions [16], political polarization [16], personality traits [46], and personality differences between democrats and republicans [52].

Given this body of research, we propose that Twitter can be used as a proxy to understand human behavior, in our case, the attitudes of Chileans regarding immigration.

3 SOCIAL THEORIES

The attitudes toward immigration are varied and depend on economic, socio-cultural and psychological factors. In this context, psychology and sociology have defined theories that explain the attitudes exhibited by people, who belong to different groups, when interacting with others: the Intergroup Contact Theory [1], and the Integrated Threat Theory [40, 49]. The attitudes toward immigration are a particular case explained by these theories.

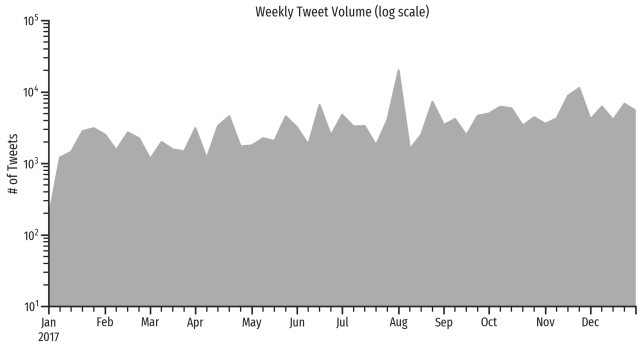


Figure 1: Weekly distribution of tweets about immigration in Chile.

Intergroup Contact Theory. Developed in the book "The Nature of Prejudice" by Gordon W. Allport [1], it postulates that prejudices are reduced when the interaction between different groups meets the following conditions: 1) groups are on equal terms; 2) they have common goals; 3) there is cooperation; and 4) there is support from formal and/or informal institutions. The theory states that intergroup contact reduces the fear and anxiety that exists when people interact with an unknown group [51], and that it promotes empathy and understanding towards the foreign group [50].

This theory has been used to ground several studies: contact between white and black people [7], heterosexuals and homosexuals [27, 28], minority religious groups [42], and locals and immigrants [30]. All these studies conclude that contact improves relationships between groups.

Integrated Threat Theory. In contrast to the Intergroup Contact Theory, the Integrated Threat Theory argues that contact between disparate groups provokes perceptions of threat and contempt [40, 49], for instance, due to competition for work and economic resources [15, 22]. Furthermore, the threat does not have to be real, it can be subjective or fictitious [33].

This theory postulates that, when the interaction conditions are not optimal, the contact between different groups will provoke conflicting and hostile relationships. The concept of "contact" is not limited only to physical contact, it can also be indirect [14], imagined [12], and electronic [2, 53].

Both theories tell us what to search when we look attitudes toward immigration: attitudes motivated by empathy, in favor of immigration; and attitudes motivated by threat, against immigration. As such, we will assume that there are two attitudes, which we label as *empathy* and *threat*.

4 DATA SET DESCRIPTION

In this section we describe our data set of posts from Twitter about migration in Chile.

Twitter is a micro-blogging platform, where users publish tweets (posts) with a maximum of 280 characters. Users may follow others,

Figure 2: Wordcloud of most frequent words in the dataset, after removing stopwords. Color is assigned according to the following categories: words, hashtags, mentions, and URLs.

to see their tweets in their own timelines. Tweets may mention other users, quote other tweets, or retweet another tweet to share it with one's audience. Users can report a screen name, a full name (which can be real or fictitious), a location (real, fictitious, or empty), and a small autobiography, among other attributes. To collect tweets that talk about immigration in Chile we used the Twitter Streaming API using system designed to crawl Chilean tweets [21]. The query parameters were keywords related to immigration (*e.g.*, inmigración, inmigrante, fronteras, racismo, *etc.*), and origin countries with their respective demonyms (*e.g.*, Haití–haitianos/as, Venezuela–venezolanos/as, Perú–peruanos/as, *etc.*). Given how generic some of these keywords are, particularly regarding the context of political issues of neighbouring countries, and the presidential elections held in Chile during November and December, we performed extensive manual clean-up of the data set.

In total, our data set is comprised by 206,353 tweets that discuss immigration in Chile during 2017, written by 49,346 users. Figure 1 shows the weekly volume of tweets. As seen on the figure, the amount of tweets has a sligth positive trend. Two peaks draw our attention: July 31th, when the news reported a case of an Haitian citizen with Leprosy; and November 19th, when an Haitian citizen rescued a woman who fell from the ninth floor of a building.

Regarding content, Figure 2 shows the most frequent words, after removing stopwords and accents. One can see that words such as Haití and Haitianos are more relevant than other countries name or demonyms, despite the fact that the largest immigrant population comes from Perú, Colombia and Venezuela [31]. Also, Santiago and Antofagasta are two frequent keywords, the two cities with the largest immigrant population [31]. Other relevant words that appear are: "gobierno" (government), "carabineros" (policemen), "Piñera" (current president, and presidential candidate in 2017) and "proyecto" (project); possibly because during the year an immigration reform was being discussed.

We explored the data set to seek for words, phrases, and hashtags that could be mapped to the *empathy* and *threat* attitudes. In *empathy* we chose terms that indicated that immigrants are welcome and will be received in equal conditions (*e.g.*, "we are all immigrants"). In *threat* we chose terms and words that showed that immigrants are not welcome and qualified them negatively (*e.g.*, "illegal immigrants"). Table 1 shows some examples of the the terms we associated to both attitudes. These labeled terms are not necessarily frequent, however, the methodology that we describe in the next sections allows to propagate these labels through a topic model.

Table 1: Examples of training terms for each attitude.

Attitude	Training Terms
Empathy	#todossomosmigrantes, #stopxenophobia, #chilesinbarreras, #chileterecibe, #bienvenidosmigrantes, @oimchile, bienvenidos a chile, #derribandomuros, @sjmchile, …
Threat	#vendepatria, #nomasinmigrantes, #nomasilegales, #inmigrantesilegales, inmigrantes delincuentes, inmigracion descontrolada, indeseables, …

5 METHODOLOGY

In this section we describe how to characterize users and tweets according to their attitude toward immigration. We define how to apply machine learning techniques to user profiles to derive user-attitude and term-attitude associations. Then, we define how to characterize attitudes from sentiment, lexical and network perspectives.

5.1 Attitudes and Topic Modeling

Topic models are a family of techniques used to discover the underlying semantic structure of a corpus by identifying and quantifying the importance of representative themes in all documents [6]. Topic models assume that each text document is generated by a set of topics which have a determined distribution. At the same time, each topic is defined by a set of words, which also have a particular distribution for each topic.

A popular topic modeling technique is Non-negative Matrix Factorization [38]. NMF works by constructing a k-rank factorization of a positive document-term matrix V into $W \times H$. Matrices W and H are estimated by minimizing the following objective function:

$$D_{NMF}(W,H) = \| V - W \times H \|_F^2, \quad W,H \geq 0, \tag{1}$$

where $\| \cdot \|_F$ is the Frobenius norm. In topic modeling, k, W and H have a special interpretation: k is the number of topics, W_{ij} quantifies the relevance of topic j in document i, and H_{ij} quantifies the relevance of term j in topic i.

Typical topic modeling applications select different numbers of k based on metrics such as perplexity. However, the meaning of topics is not always interpretable, as the factorization may follow latent patterns not necessarily aligned with human expectations. Based on the social theories described in Section 3, we propose to guide the learning procedure to seek for two topics: one that represents *empathy*, and another that represents *threat*. In such cases, supervised methods could be employed, however, these methods require a fully labeled data set, not available in our case. Since it is possible to map specific terms (words, phrases, hashtags, URLs, *etc.*) into these two topics, we propose to use a semi-supervised version of NMF known as Topic-Supervised NMF [39]. TS-NMF defines the

minimization problem as follows:

$$D_{TS}(W,H) = \| V - (W \circ L)H \|_F^2, \quad W,H \geq 0, \tag{2}$$

where $\circ$ is the Haddamard product operator, and L is a supervision matrix, defined as $L_{ij} = 1$ if topic j contributes to the document i, and $L_{ij} = 0$ if the topic j does not contribute to the document i. Thus, TS-NMF allows to provide examples of documents labeled with known topics, and to restrict the latent representation of the corpus to align with the labeled examples.

In our context, we work with user profiles, *i.e.*, the concatenation of tweets by a single user is one document. As terms we consider hashtags, mentions, URLs, and n-grams with n up to four. This allows us to define how specific phrases are mapped to each topic. The user corpus is represented as a document-term matrix D weighted with TF-IDF [4], and then row-normalized with L2 norm. To label users in the supervision matrix, we construct a list of seed terms for each theory. Then, for each row in D we estimate a preliminary attitude score for each topic, by adding the values of the cells of the corresponding seed terms. All users with a score above a certain threshold are labeled with the corresponding topic. In our experiments, we defined a threshold of 0.25, implying that only users who strongly used the seed terms of each topic were labeled.

As result, we obtain $D = U \times T$, where the rank of U and T is two. In our context, each topic is an attitude, the matrix U contains the user-attitude associations, and the matrix T contains the term-attitude associations (transposed). We interpret these associations as probabilities.

5.2 Attitude Tendency and Polarity

To characterize attitudes, we calculate two metrics common in the sentiment analysis literature to measure the leaning and amount of sentiment: tendency and polarity [35]. Tendency is defined as:

$$\text{tendency}(u) = P(\text{empathy} \mid u) - P(\text{threat} \mid u), \tag{3}$$

where, $P(\text{attitude} \mid u)$ is the association between user u and the corresponding attitude. Note that the definition is analog for terms. For tweets, tendency is defined as:

$$\text{tendency}(\text{tweet}) = \sum_{\text{term} \in \text{tweet}} \text{tendency}(\text{term}). \tag{4}$$

Note that tendency values close to zero do not imply a neutral attitude, as there could be non-zero contributions in both topics. To clarify this fact, we consider attitude polarity as the amount of associations to both attitudes, defined for users as:

$$\text{polarity}(u) = P(\text{empathy} \mid u) + P(\text{threat} \mid u). \tag{5}$$

The definition for terms is analog. For tweets, polarity is defined as:

$$\text{polarity}(\text{tweet}) = \sum_{\text{term} \in \text{tweet}} \text{polarity}(\text{term}). \tag{6}$$

In this way, tendency will allow us to group users/tweets (according to their attitude), while polarity will allow us to measure the intensity of the discussion (how polarized is the attitude).

5.3 Lexical Characterization

The previous metrics give an overview of user and tweet attitudes. The next step is to characterize grouped tweets belonging to each attitude according to their tendency. To do so, we use a psycholinguistic lexicon named "Linguistic Inquiry and Word Count" [43]. LIWC is a lexicon used to study emotional, cognitive and structural components contained in a text. In its Spanish version, it contains 7,515 words classified in one or more of 72 categories. Categories are classified into four dimensions: 1) standard linguistic processes (*e.g.*, articles, prepositions, pronouns, *etc.*); 2) psychological processes (*e.g.*, positive and negative emotions); 3) relativity (*e.g.*, time, verb tense, motion, space); and 4) personal matters (*e.g.*, sex, death, home, occupation, *etc.*). LIWC categories are organized hierarchically, for instance, all words related to the category *anger* are also organized in the categories of *negative emotions* or *affect* words.

We seek to estimate the association of tweets by tendency groups to LIWC categories. After classifying tweets into groups, we estimate how associated the words in LIWC are to each group. Note that specific events may entice a more active discussion by either group, increasing the amount of tweets, thus, we need a way to control the association with these activity patterns. In previous work, this has been done to estimate gross community metrics with z-scores [34, 45]. In our case, the definition is as follows:

$$Z_{lt'} = \frac{P_{lt'} - \mu_l}{\sigma_l}, \tag{7}$$

where, $Z_{lt'}$ is the association of LIWC category l with the tendency t', $P_{lt'}$ is the mean of fraction of words in l in each tweet with tendency t', μ_l is the mean of fraction of words in l in all tweets, and σ_l is the standard deviation of the fraction of words in l in all tweets. Hence, this relative metric allows us to compare behavior between groups, by controlling for external variability.

5.4 Network Assortativity

The previous definitions capture the behavior in expression, however, the social aspect of Twitter allows to also capture network behavior. We focus on two different networks: the mention network, related to discussion, and the retweet network, related to information diffusion. In both networks, node are users, and links are weighted relations between users. Each node has as attributes its associations to each attitude. In the mention network, a directed link between users u_1 and u_2 exists if u_1 mentions u_2 in one or more tweets. The link weight is the number of times this happens. In the retweet network, a directed link between users u_1 and u_2 exists if u_1 republishes content by u_2. The link weight is the number of times that one user retweets another. These kind of networks are commonly analyzed to understand polarization [10]. To be able to analyze connectivity, we will focus on the Largest Strongly Connected Component of each network.

To analyze the networks structure, we estimate the assortativity coefficient with respect to each attitude. The assortativity coefficient is the Pearson correlation coefficient of numerical attributes between pairs of linked nodes (this numerical attributes are the attitudes given by the model). It measures the similarity of connections in the graph with respect to the given numeric attribute [41]. Hence, the assortativity coefficient measures whether people relations are homophilic with respect to attitude. This behavior is commonly found in networks [5], and it has been documented in Twitter political discussion [10], including in Chile [20].

Figure 3: Most associated words to each attitude according to the TS-NMF model. Note that only single words are displayed, to avoid repetition in n-grams.

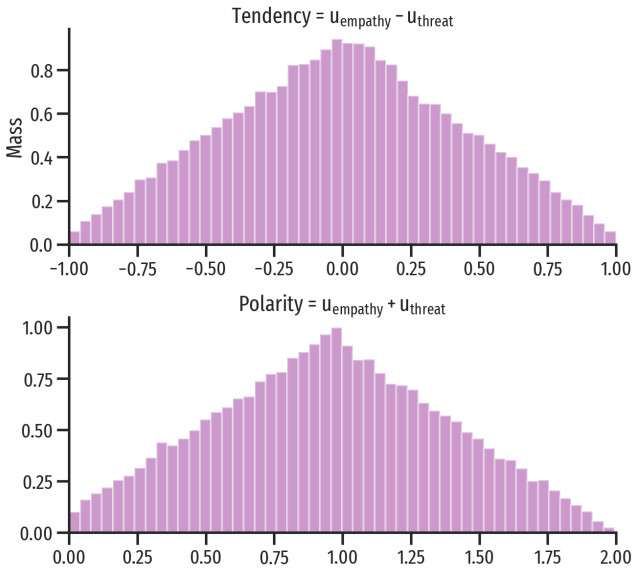


Figure 4: Top: tendency distribution for users. Bottom: polarity distribution for users.

In the next section we apply this methodology to the data set described in Section 4, covering an entire year of discussion about immigration in Chile.

6 RESULTS

Here we present the results of applying the methodology from Section 5 to the data set from Section 4.

Term Associations. Figure 3 shows the association of words with each attitude, empathy on top, threat on bottom. One can see that words associated to empathy include "integración" (integration), "salud" (health), and "educación" (education), reflecting their empathetic attitude. Words associated to threat include "delincuentes" (delinquents), "control" (control), and "ilegales" (illegals), reflecting a feeling of threat. Also, empathy group uses the word "Migrantes" (migrants) and threat group uses "Inmigrantes" (immigrants), which can be interpreted as that the empathy group is concerned about the general phenomenon (migration includes emigration and immigration), while the threat group only for the particular phenomenon (immigration).

Tendency and Polarity. Figure 4 shows the distribution of tendency and polarity for users. One can see that the distributions are fairly symmetric, with peaks in the center of the distribution. Figure 5 shows the tendency and polarity of tweets during the year under study, estimated using LOWESS. One can see that the tendency trend exhibits two interesting periods, before and after the news about the Leprosy case of an Haitian in July 31th. In the first period, tendency is slightly negative (threat), with an arguably low variability. In the second period, variability increases, and a small negative trend appears, even though at a point in time it reaches its maximum value (*i.e.*, maximum empathy) at the beginning of October. This could be explained by a news event reported in October 6th, about a Colombian citizen that gave birth on the street because a taxi driver expelled her from his car.

It is interesting that both news are related with the Integrated Threat theory and Intergroup Contact theory, respectively. On the one hand, the first event shows the immigrant as a threat, being a possible source of contagion of a disease (Leprosy). On the other hand, the second event shows the immigrant being a victim of violence and discrimination, which arguably makes people more empathetic. Regarding polarity, the trend exhibits a gradual increase in time, with two interesting peaks. The first one reflects the Leprosy case, and the second one reflects the presidential elections, where migration was a common topic in discussion.

LIWC Analysis. Figure 6 shows the differences of cognitive and emotional categories from LIWC in tweets grouped by tendency: *empathy* contains all tweets with tendency ≥ 0; *threat*, otherwise. For each category and group, we estimated the z-score for all tweets each month. As a general observation, one can see that both groups tend to have opposite behaviors. For instance, tweets in the empathy group are positively associated to the *sociability*, *family*, and *positive emotions* category more than tweets in the threat group. Conversely, tweets in the threat group are positively associated with *money*, *job*, and *inhibition* categories. This could be explained by the threat theory, as immigrants can be perceived as an economic threat and labor competition. Also, inhibition category can be interpreted by the desire to prohibit the arrival of more immigrants or to prevent them from accessing social benefits.

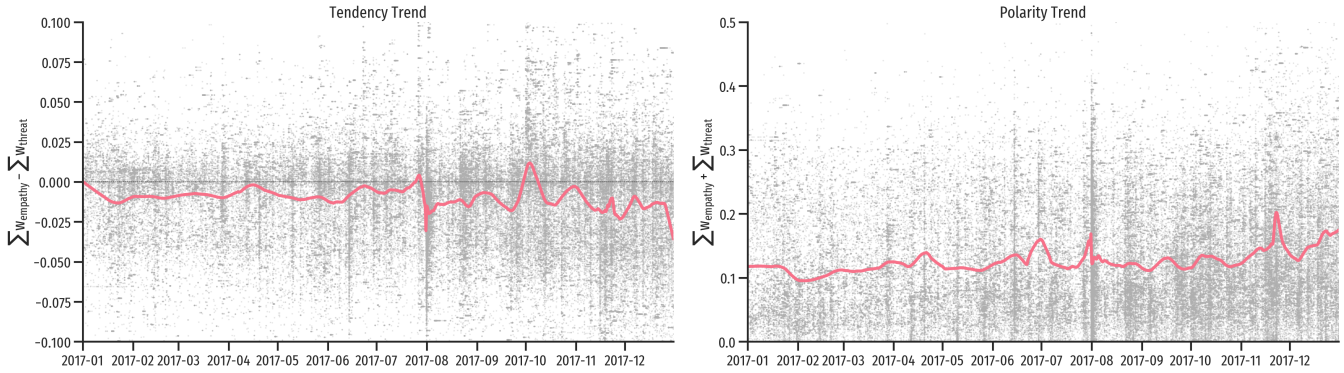


Figure 5: Trend distributions (top: tendency, bottom: polarity) for all tweets in the data set. Each tweet is a point, the x-position encodes its publication date, the y-position encodes its tendency or polarity. The line is the LOWESS interpolation of tendency and polarity.

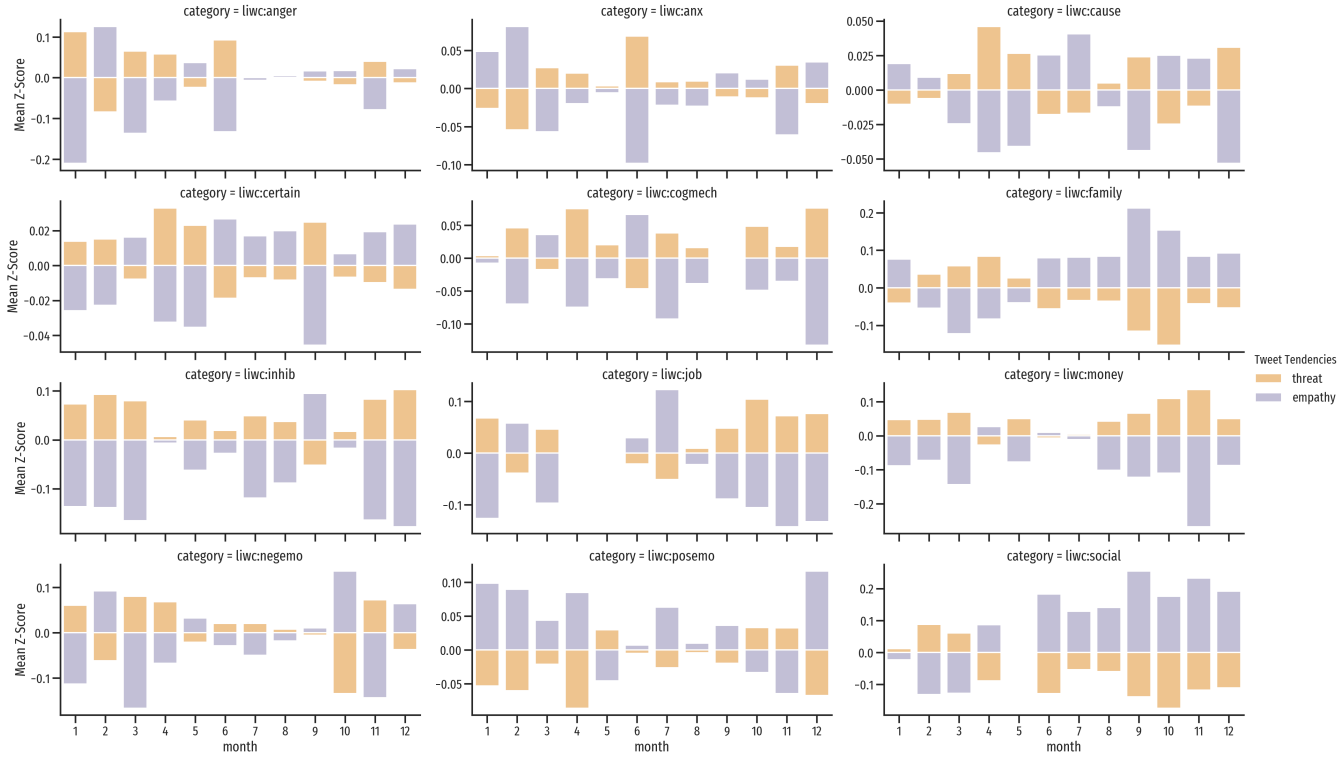


Figure 6: Association between attitudes (*empathy* and *threat*) and LIWC categories, per month. Each bar represents the association between groups, estimated with z-scores of fraction of words from each LIWC category and all other words. Purple bars indicate empathy associations, orange bars indicate threat associations.

Mention and Retweet Networks. The largest SCC of the retweet network has 1,239 nodes and 6,441 edges, while the largest SCC of the mention network has 1,868 nodes and 10,201 links. Figure 7 visualizes both networks using Hierarchical Edge Bundling [29]. This method allows us to make explicit the adjacency relations between users, as similar edges are bundled to decrease visual clutter. In the figure, each link is colored according to tendency of the source node (purple: empathy group, orange: threat group). Note that the visual encoding makes explicit the community structure in the retweet network and the heterogeneity of the mention network.

The assortativity coefficient for the retweet network are 0.26 (empathy) and 0.14 (threat), implying that homophilic behavior exists, but it is not as strong as in other topics (for instance, the discussion about abortion in Chile is greater [20]), and it is not

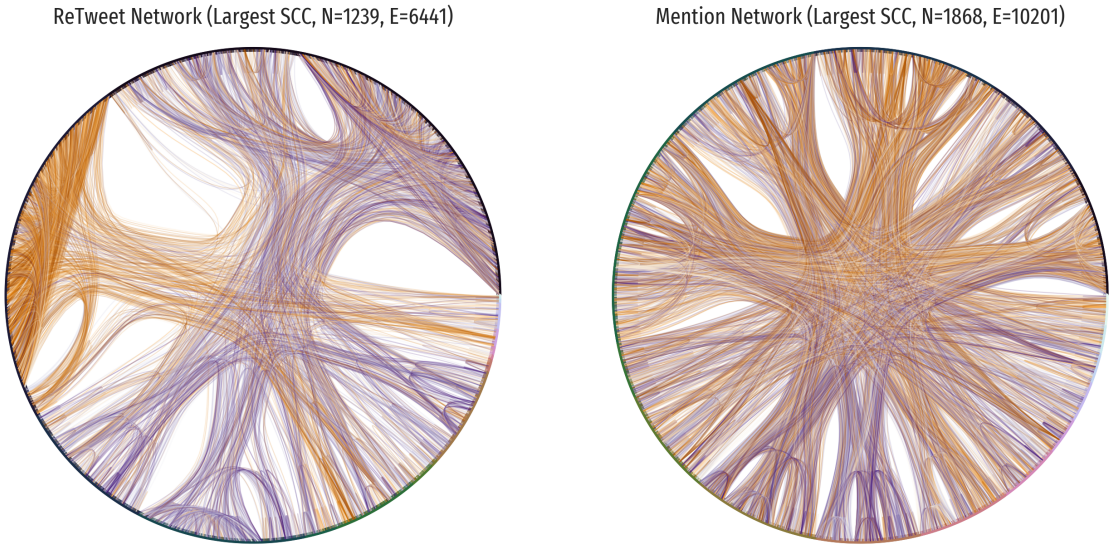


Figure 7: ReTweet Network (left) and Mention Network (right). Each node is a circle in the outside, sorted according to the connectivity patters to other nodes. Edges are lines that join nodes, where color is the attitude of the source node (purple: empathy, orange: threat). This encoding allows to group edges that are similar in terms of connectivity between groups.

equal in both groups. As hinted by the visualization, in the mention network the results are small: 0.06 (empathy) and 0.08 (threat). Thus, the retweet network is more segregated than the mention network. This could be explained because retweets are expected to be seen by all followers, and are a key factor in information diffusion, while mentions and replies are not. For instance, one user may send tweets to another holding an opposite position, but if there is no reply, then the interaction is not meaningful.

7 DISCUSSION AND FUTURE WORK

Migration is a controversial issue in Chile, and, although there are some studies about Chileans attitudes toward immigration [9, 37], they do not cover recent migration patterns. To complement knowledge about this topic, we defined a way to classify and measure attitudes, enabling to study the dynamics of perception with respect to immigration and performed a descriptive study of how immigration is perceived in Chile, according to Twitter discussion.

Our results may inform policy and intervention design, as it quantifies how people feel and communicate with respect to immigration. This is relevant, as there exists several contact strategies to improve relationships between social groups [44]. For instance, the discussion we analyzed is mostly targeted at Haitian migration. A majority of them is from Afro-Haitian descent, an ethnicity that was almost non-present in Chile.

There are two key aspects that need further exploration, and that limit the scope of our results: the representativity of Twitter, and the validation of the TS-NMF model. In terms of representativity, Twitter is a biased sample of the population [3]. As such, our results only cover this sample, even though it is not know to which degree nor to which sub-populations it represents. Having these biases into account will surely improve the interpretation of results. However, one aspect that needs to be considered is that Twitter is within the most popular applications in Chile [18], and that it reflects some cultural aspects, such as the country's centralization [19]. In terms of validation, the lack of ground truth or approximate measures of the problem stands in the way of effectively measuring the model accuracy, leaving us only with a qualitative evaluation.

Besides working on the limitations of our approach, there are two lines of future work that we devise. On the one hand, it would be relevant to understand the relationship between attitudes and actual presence of immigrants in a place. This would provide a way to measure real and imagined threat attitudes [33]. On the other hand, there is a potential influence of news events in attitudes. Given the rise of *fake news* and *post-truth* media, this would provide a way to measure the effect of such phenomena on how people feel with respect to a specific issue, migration in this case.

8 CONCLUSIONS

In this paper, we have characterized attitudes toward immigration by locals in Chile. We used a semi-supervised topic modeling technique (TS-NMF [39]) to identify attitudes grounded in two social theories, the Intergroup Contact Theory [1], and the Integrated Threat Theory [40, 49]. Then, we measured differences in attitudes using psycho-linguistic lexicons and interaction networks. As result, we found consistent behaviour with respect to social theory. There is still work to do in the evaluation and representativeness of our model, including the definition of a suitable ground-truth perception to validate our proposal. We believe our results help to inform the design of public policy and interventions to improve relations between groups in a country.

REFERENCES

[1] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice. (1954).

[2] Yair Amichai-Hamburger and Katelyn YA McKenna. 2006. The contact hypothesis reconsidered: Interacting via the Internet. *Journal of Computer-Mediated Communication* 11, 3 (2006), 825–843.

[3] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval: the concepts and technology behind search, 2nd. Edition.* Addison-Wesley, Pearson.

[5] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23, 1 (2015), 76–91.

[6] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[7] Kendrick T Brown, Tony N Brown, James S Jackson, Robert M Sellers, and Warde J Manuel. 2003. Teammates on and off the Field? Contact with black teammates and the racial attitudes of white student athletes 1. *Journal of applied social psychology* 33, 7 (2003), 1379–1403.

[8] Peter Burns and James G Gimpel. 2000. Economic insecurity, prejudicial stereotypes, and public opinion on immigration policy. *Political science quarterly* 115, 2 (2000), 201–225.

[9] Héctor Carvacho. 2010. Ideological configurations and prediction of attitudes toward immigrants in Chile and Germany. *International Journal of Conflict and Violence (IJCV)* 4, 2 (2010), 220–233.

[10] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. *Proc. of ICWSM* 133 (2011), 89–96.

[11] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.

[12] Richard J Crisp and Rhiannon N Turner. 2009. Can imagined interactions produce positive perceptions?: Reducing prejudice through simulated social contact. *American psychologist* 64, 4 (2009), 231.

[13] Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, and Norah Abokhodair. 2017. Predicting Online Islamophopic Behavior after# ParisAttacks. *The Journal of Web Science* 3, 1 (2017).

[14] John F Dovidio, Anja Eller, and Miles Hewstone. 2011. Improving intergroup relations through direct, extended and other forms of indirect contact. *Group processes & intergroup relations* 14, 2 (2011), 147–160.

[15] Victoria M Esses, John F Dovidio, Lynne M Jackson, and Tamara L Armstrong. 2001. The immigration dilemma: The role of perceived group competition, ethnic prejudice, and national identity. *Journal of Social issues* 57, 3 (2001), 389–412.

[16] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural dimensions in Twitter: Time, individualism and power. *Proc. of ICWSM* 13 (2013).

[17] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 581–586.

[18] Eduardo Graells-Garrido, Diego Caro, Omar Miranda, Rossano Schifanella, and Oscar F Peredo. 2018. The WWW (and an H) of Mobile Application Usage in the City: The What, Where, When, and How. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1221–1229.

[19] Eduardo Graells-Garrido and Mounia Lalmas. 2014. Balancing diversity to counter-measure geographical centralization in microblogging platforms. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 231–236.

[20] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2015. Finding intermediary topics between people of opposing views: a case study. In *Social Personalisation & Search*, Christoph Trattner, Denis Parra, Peter Brusilovsky, and Leandro Balby Marinho (Eds.). CEUR, Santiago, Chile.

[21] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2016. Encouraging diversity-and representation-awareness in geographically centralized content. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 7–18.

[22] Shang E Ha. 2010. The consequences of multiracial contexts on public attitudes toward immigration. *Political Research Quarterly* 63, 1 (2010), 29–42.

[23] Jens Hainmueller and Michael J Hiscox. 2007. Educated preferences: Explaining attitudes toward immigration in Europe. *International organization* 61, 2 (2007), 399–442.

[24] Jens Hainmueller and Michael J Hiscox. 2010. Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment. *American political science review* 104, 1 (2010), 61–84.

[25] Gordon H Hanson, Kenneth Scheve, and Matthew J Slaughter. 2007. Public finance and individual preferences over globalization strategies. *Economics & Politics* 19, 1 (2007), 1–33.

[26] GACCT Harman and Mark H Dredze. 2014. Measuring post traumatic stress disorder in Twitter. *In ICWSM* (2014).

[27] Gregory M Herek. 1986. The instrumentality of attitudes: Toward a neofunctional theory. *Journal of Social Issues* 42, 2 (1986), 99–114.

[28] Gregory M Herek and John P Capitanio. 1996. "Some of my best friends" Intergroup Contact, concealable stigma, and heterosexuals' attitudes toward gay men and lesbians. *Personality and social psychology bulletin* 22, 4 (1996), 412–424.

[29] Danny Holten. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics* 12, 5 (2006), 741–748.

[30] Daniel J Hopkins. 2010. Politicized places: Explaining where and when immigrants provoke local opposition. *American political science review* 104, 1 (2010), 40–60.

[31] National Statistics Institute (INE). 2018. Síntesis resultados Censo 2017. urlhttps://www.censo2017.cl/descargas/home/sintesis-de-resultados-censo2017.pdf.

[32] Seth K Jolly and Gerald M DiGiusto. 2014. Xenophobia and immigrant contact: French public attitudes toward immigration. *The Social Science Journal* 51, 3 (2014), 464–473.

[33] Jeffrey S Kopstein and Jason Wittenberg. 2009. Does familiarity breed contempt? Inter-ethnic contact and support for illiberal parties. *The Journal of Politics* 71, 2 (2009), 414–428.

[34] Adam DI Kramer. 2010. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 287–290.

[35] Onur Kucuktunc, B Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. 2012. A large-scale sentiment analysis for Yahoo! answers. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 633–642.

[36] Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J Ramasco. 2018. Immigrant community integration in world cities. *PloS one* 13, 3 (2018), e0191612.

[37] Duncan Lawrence. 2015. Crossing the Cordillera: immigrant attributes and Chilean attitudes. *Latin American Research Review* 50, 4 (2015), 154–177.

[38] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788.

[39] Kelsey MacMillan and James D Wilson. 2017. Topic supervised non-negative matrix factorization. *arXiv preprint arXiv:1706.05084* (2017).

[40] Todd D Nelson. 2009. *Handbook of prejudice, stereotyping, and discrimination*. Psychology Press.

[41] Mark EJ Newman. 2003. Mixing patterns in networks. *Physical Review E* 67, 2 (2003), 026126.

[42] Josef Novotny and Filip Polonsky. 2011. The Level of Knowledge about Islam and Perception of Islam among Czech and Slovak University Students: does Ignorance Determine Subjective Attitudes? *Sociologia* 43, 6 (2011), 674–696.

[43] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[44] Thomas F Pettigrew and Linda R Tropp. 2006. A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology* 90, 5 (2006), 751.

[45] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 965–968.

[46] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our Twitter profiles, our selves: Predicting personality with Twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 180–185.

[47] Kenneth F Scheve and Matthew J Slaughter. 2001. Labor market competition and individual preferences over immigration policy. *Review of Economics and Statistics* 83, 1 (2001), 133–145.

[48] Paul M Sniderman, Pierangelo Peri, Rui JP de Figueiredo Jr, and Thomas L Piazza. 2000. The Outsider: Politics and Prejudice in Italy.

[49] Cookie White Stephan and Walter S Stephan. 2013. An integrated threat theory of prejudice. In *Reducing prejudice and discrimination*. Psychology Press, 33–56.

[50] Walter G Stephan and Krystina Finlay. 1999. The role of empathy in improving intergroup relations. *Journal of Social issues* 55, 4 (1999), 729–743.

[51] Walter G Stephan and Cookie White Stephan. 1985. Intergroup anxiety. *Journal of social issues* 41, 3 (1985), 157–175.

[52] Karolina Sylwester and Matthew Purver. 2015. Twitter language use reflects psychological differences between democrats and republicans. *PloS one* 10, 9 (2015), e0137422.

[53] Fiona A White and Hisham M Abu-Rayya. 2012. A dual identity-electronic contact (DIEC) experiment promoting short-and long-term intergroup harmony. *Journal of Experimental Social Psychology* 48, 3 (2012), 597–608.

#Default #Interactiveart #Audiencexperience

Learning from the visual analysis of content shared on Instagram

Manuela Garretón
Departamento de Ciencias de la Computación
Pontificia Universidad Católica de Chile, Santiago, Chile, mgarret1@uc.cl

Andrea Rihm
Laboratorio de Investigación en Subjetividad y Cambio Social
Santiago, Chile, arihm@uc.cl

ABSTRACT

The last decades have shown us a growing interest in different fields of how interactive art transforms the position of the viewer into a participant and how audiences engage and relate to interactive artwork. This article presents a visual analysis of the content shared on Instagram by the audience of Default –an interactive art installation presented in Santiago, Chile in 2017. The analysis shows that people reacted and engaged differently with various aspects of the installation, as shown by the strategies they used to share it. We argue that the analysis of the visual content of Instagram posts opens avenues to understanding the relationship between installation and audience, giving clues on the audience experience and, therefore, providing feedback for developers, who could use them in the design process of future installations.

CCS CONCEPTS

• HCI design and evaluation methods • Social media networks • Media arts

KEYWORDS

Interactive art evaluation, Instagram, Social Media, Arts, and culture on the web

ACM Reference format:

Manuela Garretón and Andrea Rihm. 2019. **#Default #Interactiveart #Audiencexperience:** Learning from the visual analysis of content shared on Instagram. In *Proceedings of WWW '19: The Web Conference (WWW '19). ACM, New York, NY, USA,* 8 pages. https://doi.org/10.1145/3308560.3316453

1 Introduction

The use of digital technologies in art has transformed the visitor into a participant, where the visitor no longer contemplates an object or artefact, but has access to an experience [9]. This has led many artists to include evaluations in their work in order to understand which aspects of an art installation impact the user's experience and perception [9,38,4]. The approximations to evaluate art installations vary from documented reflective practice to evidence-based methods [11]. Much of the progress in this context can benefit from digital art and HCI research. Nevertheless, in interactive art the interest generally focuses on how the artwork behaves, how the audience interacts with it and their degree of engagement [16].

In this article we focus on the latter, using Instagram as a way to analyze the experience of the audience while visiting an interactive art installation and learn from this to enhance future iterations of the artwork. *Default*, the installation that provided context for this work was exhibited at the Contemporary Art Museum of Santiago de Chile in 2017. It explored the dynamics of consciousness and memory through visualization and sonification of data. It is inspired by the concept of the neuroscience discovery of the default mode network (DMN).

While the installation was exhibited, visitors spontaneously shared images and videos on Instagram about it. This meant –on the one hand– that the interaction with the installation was taken beyond the museum walls and into the digital space, therefore amplifying its reach and impact and –on the other– that the audience appropriated the installation, reframing it and creating new artwork based on their experience.

By observing this phenomenon, we formulated the following questions: Could instagram posts be used as data to analyze the visitors' experiences? Could they be visually analyzed, in terms of their content, to better understand the impact of the installation? How could these data be used to learn and enhance future design processes?

Following these questions, this paper presents and discusses the results of the analysis of the visual content of Instagram posts[1], which proved to be useful in accessing the audience experiences, particularly with regard to the interactive, immersive and aesthetic qualities of the installation. Despite the fact that this work is based on a relatively small number of cases (53), we found evidence to argue that this type of analysis provides useful feedback from the audience experiences to the researchers, regarding what captivated them and how they interacted with the installation. We discuss that these data that can be incorporated in future iterations and design processes, providing guidelines for eventual artistic installations. Given

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW '19, 2019-05-13, 2019, San Francisco, USA
© *2019 Association for Computing Machinery.*
ACM ISBN 978-1-4503-6675-5...$15.00
https://doi.org/10.1145/3308560.3316453

[1] From now on, "Instagram post" will be used to refer to the visual content published, not including captions, hashtags, comments or likes of the post.

that social media analysis tends to use quantitative data to assess the impact and engagement with the content share, we argue that the visual analysis of the visual content also contributes to enrich such analysis by providing alternative avenues to access qualitative data of the audience experiences, particularly for audiences of 35 years old or younger, who constitute the larger group of Instagram users.

2 Context

Default was an interactive artistic installation that explored the dynamics of consciousness and memory through visualization and sonification of data, inspired by the concept of the neuroscience discovery of the default mode network (DMN) [30]. The DMN consists of a group of brain regions, which activate when the mind starts wandering. The project was led by one of the authors of this article: Manuela Garretón (design) in collaboration with Tomás Ossandón (cognitive neuroscience).

Default consisted of the representation of the two main brain states, the resting state (when the DMN is active and the brain registers a high level of activity) and the attention state (when the brain responds to an external stimulus and the DMN is disabled). The installation was conceived as a visual and sound environment based on data recorded in a functional magnetic resonance imaging (fMRI) of a person in resting state. The interior space of the room was designed as if it were the interior of the brain [17]. The visitors acted as an external stimulus, activating the attention state. Default was designed to generate a visual, aesthetic and immersive environment where the audience not only sees the visualization, but can also enter, move and navigate through it. A 3D animated visualization was designed to be projected in three continuous walls of the room. The system (visualization and sonification) responded to the behavior of the audience through the implementation of two inputs. An infrared camera (Kinect), identified people and their positions inside the room, and reacted to their movements changing the projections up the wall, therefore, creating an interactive loop.

Default was presented in the Museum of Contemporary Art (Parque Forestal) in Santiago, Chile in 2017 . The goals of the installation were: (1) To display the representation of intrinsic dynamics of brain in an interactive artistic space (2) to create a medium to disseminate the existence of the DMN and its related concepts for a non-expert audience (3) To captivate and engage the audience with an aesthetic, interactive and immersive experience. The objectives of this project defined three complementary research line and this article focuses on the one related to the understanding of the audience experience in Default.

By design, the audience experience in *Default* was supposed to be assessed only through videos registered by the research team. However, during the exhibition, it became apparent that some visitors were spontaneously uploading and sharing their experiences on Instagram, therefore making an explicit gesture of appropriation of the installation. In reviewing these posts we wondered about the great potential for analysis that this material could have to access the experience of visitors in Default. This was an emergent result of the exhibition that gave us a privileged glimpse into the audience experience and what was interesting to them. The work that follows describes the questions that arose from the first observations of this collected data and the subsequent analysis.

3 Related Work

3.1 From Neuroscience to an interactive artwork

In order to understand the questions we asked ourselves about the visitors' experience it is important to underscore two defining features of Default as an art installation: its interactive nature and its roots in scientific concepts. Both are key to situate Default in a multidisciplinary field.

In the last decades, the combination between arts, sciences and technologies, including new media, digital art and interactive art [38] has brought knowledge to non-expert audience, offering a unique contribution to engage the public with complex scientific issues [24]. Under the concept of Sci-Art, a new approach has developed to engage broad audiences with complex scientific knowledge, by exploring a new language that emerges from intimate, sensory, persona, human-scale narratives, metaphors and aesthetics [8]. These projects have reached various subjects such as genetics, neuroscience, climate change, astronomy, and so on, and have emerged from the traditional performance arts, interactive installations, sound art, data visualization, interactive web documentaries, among others [39,24,23,8,19,28,18].

The emergence of sci-art as a bridge between the scientific community and non-expert audiences is related with the understanding that art flourishes most when it serves a social function [5, 25]. Artwork becomes meaningful through interactions and engagements with an audience.

It is in the interaction between the visitor and the installation where the artwork completes itself –even if temporarily and in a relation-specific way– and acquires meaning. The characteristics of the interaction are shaped by the singularities of each visitor and to know them would be crucial to developing future iterations of the artwork. Consequently, it becomes important to have access to that personal and situated experience. The question remained is how to do it.

Given all the above, it was possible to conceive Default as an interactive sci-art installation. This conception, however, required a thorough analysis of the experiences. Had they not interacted with the installation, this intended feature would not have come to fruition. Yet, how could we learn about the visitors' experiences? How can we learn about their interaction with Default? Instagram as a visual social media proved to offer an insightful glimpse of the visitors' experiences.

3.2 Instagram as a platform to share experiences

Instagram is a social network that allows registered users to upload pictures and videos to a mobile and a desktop application. Users have the capability to caption the content, add locations through geotags, hashtag their posts and tag other users. By now (January 2019) it has reached over one billion monthly active users [22]. Of the total users, 68% are younger than 35 years old [35]. Even though until now this social network has not been vastly used, it is a very useful tool to obtain data, images, video and opinions regarding social research. It has been used for the analysis of user experiences in various areas such as cultural studies [21,26], urbanism [36] and more specifically to discuss the engagement of exhibitions at a museum [37].

The predominant visual language of Instagram is relevant given that the production of images has been considered a meaningful way to express one's experiences in a language that has gained greater centrality in our culture, that plays a major role in most people everyday life [31,32,33,43] and that allows for expression and articulation of aspects and experiences that are not always easy to verbally express. Moreover, it has been argued that the communal photographic exchange occurring over Instagram is implicated in the production of identity, as cultural consumers and producers [8].

On the other hand, the understanding of art as an experience has lead digital artists to include evaluations within their works with several purposes: mostly, to understand how the artwork is perceived by the audience and get a broad view of the visitor experience [10,11,6,7]. The approximations to evaluate interactive art vary from documented reflective practice to evidence-based methods [10, 11]. Much of the progress in this context can benefit from the digital art and HCI research. There is constant feedback between both areas. Many of the methods practiced in HCI are used as starting points to evaluate artistic experiences. Some of these go from quantitative methods to identifying patterns of interaction [18] and qualitative analysis based on observations between the visitor and an interactive artwork [6]. Nevertheless, in interactive art the interest generally focuses on "how the artwork behaves, how the audience interacts with it and, (…) ultimately, in participant experience and degree of engagement" [16].

In the case of Default, once we realized that visitors were sharing Instagram posts about their visit we saw an opportunity to approach the evaluation of their experience within the artwork. We observed that visitors not only were sharing part of their experiences on a visual way, but they were also appropriating the installation, re-framing it from their own perspective and, by doing this, somehow creating a new artwork. For us this new artworks become a way to have access to that personal experience. So, what can we learn from the images created from the users' experience in an interactive space? Could they be a way to approach the evaluation of interactive art experiences?

4 Materials and Methods

This article builds on the analysis of emerging results of the interactive art installation, more than corresponding to an independent research project. However, it could be conceptualized in the line of what has been defined as arts-informed research [12, 43], which has among its goals, accessibility and breadth of audience and which is "part of a broader commitment to shift the dominant paradigmatic view that keeps the academy and community separated, to acknowledge the multiple dimensions that constitute and form the human condition […] and the myriad of ways of engaging in the world—oral, literal, visual, embodied" [12]. Therefore, both the arts-informed research audience engagement and a social commitment are both key.

4.1 Data Handling

To carry out this study we searched for every Instagram public post made by the audience while the exhibition was open to the public (8th September 2017 - 12th November 2017). We searched by location in the Instagram website using the geotag of the museum "MAC / Museo de Arte Contemporáneo". Also, we searched for the National Fine Arts Museum geotag "Museo Nacional de Bellas Artes - Santiago de Chile". Because both museums are located in the same building and we corroborated that several users might have made this mistake when selecting from the prompted list that Instagram provides based on the device location. We found a total of 53 posts that belonged to publicly accessible Instagram user accounts. Once we identified all the posts we firstly collected their urls and then used import.io, a free online tool that allows people to extract the items on a website into a dataset. We gathered images and video urls, names of user, captions, number of likes, among others. Next, we downloaded all the images and videos so that they could be maintained for exploration during analysis.

4.2 Data Analysis

Previous works have highlighted that one of the main challenges in the use of images or videos for research is the lack of a standard "set of rules" to analyze them, given the polysemic nature of visual data [20, 34, 40]. For the purpose of this research, we analyzed all the available Instagram posts (53) following proposals by Acosta [1] Rose [34] and Gleeson [20] in terms of what to focus on and how to conduct the analytical process.

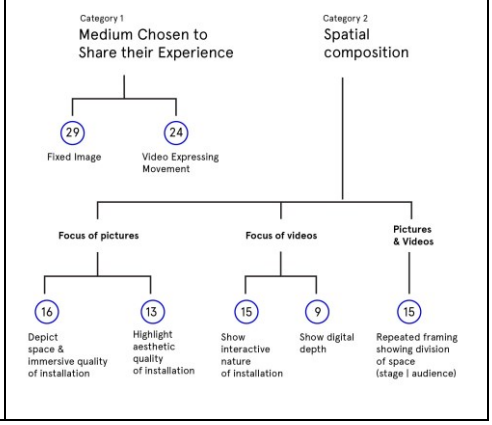
Acosta [1] highlighted the importance of observing and analyzing the aesthetic and dynamic properties of each image. Rose [34] focused her visual analysis in examining the content of images to identify patterns, coding and analyzing such patterns to draw conclusions, and Gleeson [20] in her polytextual thematic analysis for visual data underscores the importance of analyzing images in a number of iterations, to capture proto-themes that eventually turn into higher-order categories.

Following the above mentioned methodologies, and after the first iteration analyzing our data, it became evident that a number of criteria were relevant to systematically observe in order to understand the images and videos (Fig. 1). Following the emerging criteria, the analytical process was iterated a number of times with images and videos projected on a 51-inch screen.

Each picture was first analyzed individually, secondly, all the pictures were analyzed together, with a focus on recurring patterns, themes and differences, using the criteria shown in Figure 2: (1) medium chosen to share the experience, (2) aesthetic qualities of the image and (3) site of production of the

image. Then, videos were analyzed following the same steps. Finally, the joint data from videos and pictures were analyzed. From each analytic stage a number of proto-themes related with the experience of the visitors were developed. Eventually, these proto-themes were grouped to develop two higher order categories that connected all of them, and that serve to identify different appropriation strategies of the installation utilized by the visitors. The analytic process, in all its iterations, was conducted by two researchers as a way to increase the rigor, and reliability of the process.

MEDIUM CHOSEN TO SHARE EXPERIENCE	Video
	Picture
AESTHETIC QUALITIES OF THE IMAGE	Spatial organization - Composition *- Focus of the image/video* - Vantage point *- Close up shots* *- Distance shots*
	Presence or absence of people -Facing the camera - Facing sideways - Giving their back to the camera
	Presence or Representation of movement - Type of movement *- Movement of people* *- Movement of projections on the wall* *- Movement associated to depth*
SITE OF PRODUCTION OF THE IMAGE	From the "stage" space
	From the audience space



Figures 1 and 2: (Left) Emerging relevant criteria used to analyze Instagram posts (Right) Strategies used to share experience, and number of appearances (N=53)

5 Findings

It is possible to assert that a number of people (53) felt captivated enough by Default to spontaneously register and publicly upload an image or video about it. In addition, a significant number of followers felt something similar, to like those posts (1,296) and view the videos (2,054). This is relevant, given that it was not intentionally prompted –or expected– by the researcher's team by providing a hashtag to identify it. By tagging themselves in the museum, and sharing their experiences, people are inscribing the installation in their trajectories, therefore, appropriating it. Through their Instagram posts, people showed what was particularly interesting and/or curiosity triggering for them, offering a glimpse into to their experiences.

The analysis of the images and videos posted on Instagram allowed us to distinguish two main categories, related with strategies used by the audience to respond to the aesthetic, interactive and immersive nature of the installation: (1) Medium chosen to share their experience and (2) Spatial Composition. The results show that these two strategies do not appear in isolation, but tend to interact to some extent. (Fig. 2)

5.1 Medium chosen to share the experience

We observed that people used the two mediums offered by Instagram to share their experience: images (29) and short videos (24). The difference between mediums is meaningful given that, as Moon [29] stated, this election is one of the first forms to appropriate of the artistic/aesthetic experience. By its very nature, photographs fixate an image, while videos are suited to express movement. Thus, the chosen medium relates to two distinct modes to represent the experience within the installation: one that focuses on showing the interaction with it, moving along with it and another, that responds more to the aesthetic quality of the installation, registering it in a fixed way.

5.2 Sharing the interaction through videos

Twenty-four people utilized videos to share their experience (17 short videos, 2 videos edited with timelapse and 5 boomerang). As expected given the nature of the medium, movement was the most significant feature of these Instagram posts. However, it is relevant to note that the movement depicted was not a feature "added" by the people, but depicted the interaction with the installation, as movement is a key feature of it.

Movement was depicted through three different strategies: (1)Moving the camera across the space to show the movement of the projections (video 1)[2]; (2) Showing people moving around the exhibition; (3)Keeping the camera fixed, zooming in and out the lens to show the depth of the projections (video 2)[3].

Despite the different techniques utilized by the audience, one thing was clear: they accounted for the experience of interaction. People followed the pace and rhythm suggested by the projections, some even danced

[2] Video 1: To see an example of fixed camera and people moving visit https://www.instagram.com/p/BZ2BSXsj1vh/?taken-by=pipexs_jt [29]
[3] Video 2: To see an example of digital depth https://www.instagram.com/p/BZrGJffj1U2/?taken-by=diegorubio83 [15]

along (video 3)[4]. In doing this, people displayed their being affected by the nature of the installation and also, showed a playful behavior and sense of enjoyment.

5.3 Sharing the aesthetic quality of the installation through pictures:

The pictures shared emphasize the graphic nature of the projections, as shown by their careful framing and composition. All images focus on the projections up the wall, as shown by the fact that they occupy approximately 80% of the space (or more) of each picture. In fact, a person even edited the image (Fig. 4), drawing figures like the ones projected, to complete the space. The edition of images in terms of color (black and white; sharpening) was used to emphasize contrast and better outline figures, accentuating clarity and definition of the image. Finally, the care for composition can be observed in the balanced nature of the images: most people clearly established a center, depicting an intention to articulate the image as a whole.

5.4 Spatial Composition

The visitors responded to the space of the interactive installation differently, as shown by the spatial composition of the Instagram post, which is expressive of a perspective and a point of view, particular to the viewer. This interacted with the medium chosen by the visitors to share their experience.

5.5 Pictures:

Pictures had two main foci: (1) To depict the space and the immersive quality of the installation; (2) to highlight the aesthetic quality of the installation in relation with people's presence within the space. This difference among the images signals two different strategies of appropriation of what the installation has to offer to the audience. However, as it will be shown, no strategy is absolute, displaying some mixed qualities despite the predominance of a certain focus.

The first strategy, used by 16 people, entails photographing the space from some distance as a way to "make it appear". Some people included two angles of the room (i.e. three walls) (Fig. 3), and some only one angle (i.e. two walls), but all the pictures of this kind keenly convey the immersive quality of the installation. The person is *in* it. The composition proportion person: space reveals that the person cannot capture all what is happening on a glimpse. He/She needs to move around to interact with the space. In this sense, the installation challenges the experience of being a "mere observer" of the artwork and calls for movement. It is interesting to note that in most pictures, even though people were clearly outsized by the installation, they did not entirely relinquish a central role in the picture, tending to place themselves near the center. Also, because they appear darker than the projections on the wall, the contrast highlights their presence, therefore gaining a protagonist stance.

The second strategy entails photographing the walls from a much closer vantage point, in the mode of portraits focusing on the subject. As a consequence of this focus, the composition of the image results in a "flattening" of the space, making it appear mostly as background (figs. 4 and 5). This, however, does not mean that people did not take into account what the installation offered, or that they merely used in as a "prop" for highlighting themselves in a way that does not recognize the installation qualities. On the contrary, each of these posts –in its composition– responded and was highly attuned to the aesthetic quality of what the installation presented, therefore demonstrating the impact it had had on the people. For those who flattened the image, it could be argued that the graphic nature of the projection was more appealing than the immersive and/or interactive quality of the installation.

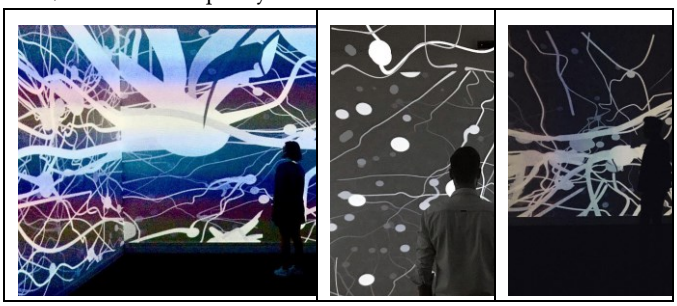


Figure 3 ,4 and 5: (Left) Person immersed in the space [13] (Center and Right) The wall projections as portrait background [28, 13]

In fact, it is interesting to consider that even though these 13 images could be conceptualized as portraits, only three people were facing the camera (and the intended viewer of the image). Most of them were either standing sideways (4) or had their back to the camera (4), all facing the projection on the wall. In two cases, the minimum light made it difficult to recognize their orientation. Therefore, it is possible to argue that if the camera focuses on the person, and the person mostly focuses on the projection, the focus of the image also shifts to the wall, emphasizing its importance.

Some "mixed cases" were those of Instagram posts made with *Boomerang*. Here the movement corresponds to only one second looping video made with ten images stitched together that look more as an animated GIF than a proper video. Therefore, *boomerangs* (5) could be

[4] Video 3: To see people dancing visit https://www.instagram.com/p/BabXYlLnrNf/?taken-by=bryanmerino.l [2]

considered as representing an intermediate position, where movement is not necessarily related with the interactive nature of the installation.

5.6 Videos

15 out of the 24 videos showed people. However, the foci of those videos were not people in themselves, but demonstrating the interactive nature of the installation, showing how people's movements –captured by the infrared camera– changed the projections on the wall, creating an interactive loop where people responded with their movements to the rhythm suggested by the projections and, in turn, the projections were affected by the people's movements (video 1).

Videos without people, where the camera moves to show a close up (video 2), focus on capturing the digital depth of the projection, showing the intention to "enter" the projected space, move through it and capture the experience of being "inside the visualization". Thus, these kinds of videos convey the immersive quality of the installation beyond the mere representation of people being *in* the space (as was shown in the pictures). They "enter" a digital space of the projected visualization that amplifies the material conditions of the real space, which is quite modest.

5.7 Pictures and videos

On the other hand, when analyzing the viewer's point of view in both videos and images, we observed significant similarities in terms of revealing the position of the person who took the picture or video. In this way we deduce that almost every picture was taken from the back of the room. Thus, what varies in every Instagram post is the depth of the registry.

Based on these observations we can conclude that the framing is similar in almost every image, the contents are essentially the same (the projections with a person facing backwards) and the element that varies upon the pictures is the depth of registration. This could be attributed to the material conditions of the room where Default was installed, which favored the creation of an unintended "virtual space". This virtual division created a "stage", which contributed to determine different places and roles inside the room for the audience. The spectator-performer who uses "the stage", interacts with the projections and sound can be recognized as a performer, and the spectator-audience –who observes from the back of the room assuming a role of an audience and captures videos or pictures of what is happening on "the stage"–.

To summarize (Fig. 2) by analyzing the images and videos in the categories we can say that:

(1) To choose images as a medium to share experiences demonstrates the focus on the *aesthetic* nature of Default, while choosing videos reveals the importance that visitors give to the *interactive* feature.
(2) The spatial composition of images shows a first group (slightly predominant) depicting space thus capturing the *immersive* quality of the installation. While a second group flattened the images, appropriating the *aesthetic* nature of Default.
(3) When people are included in videos is mostly to express the *interactive* quality of the installation. Videos without people, on the other hand, emphasize the *immersive* quality of the projected visualization.
(4) The framing of the images and videos repeats among the retrieved Instagram posts, but the depth from where they are taken varies.
(5) It is possible to recognize two types of spectators that occupy different places and roles inside the room; the spectator-performer and the spectator-audience. These roles also determine an unintended virtual space: the 'stage' and the 'audience'.

6 Conclusions

The aim of this article was to assess if the visual content of Instagram posts could be used as data to analyze the visitors' experiences, and if this content analysis of visual data could be used to better understand the impact of the installation and, how could this data be used to learn and enhance future design processes.

The findings offer evidence that is possible to visually analyze Instagram posts in terms of their visual content. The images and videos proved to be useful in distinguishing the way audiences experienced the installation, offering qualitative data regarding its impact. This offers a complement for more traditional social network analysis, which tends to provide data regarding the quantitative impact and engagement of the content shared.

In the past, analysis of Instagram posts has been used to assess visitors' experiences in museums [9, 37]. *Default*, however, had the particularity of being an interactive art installation. Our findings allow us to conclude that the visual content shared on Instagram accounts for differences between those visitors who felt more captivated by the aesthetic quality of the installation and those who highlighted the interactive and immersive quality of the installation. The fact that Instagram –as a platform– offers the possibility to share pictures and videos proved to be key in this respect. Those who felt particularly captivated by the aesthetic quality of the

installation tended to use pictures and highlight the graphic nature of the projections, sometimes at the expense of depicting the space as a whole, therefore leaving aside the immersive and interactive nature of the installation. People who felt mostly captivated by the interactive nature of the installation privileged the use of videos that allowed conveying movement, and how their movements and the movements of the projection mutually influenced each other. Finally some people were more responsive to the immersive quality of the installation, as shown by their representation –through pictures– of being *in* the space and –through videos– to enter the digital depth of the projection, virtually amplifying the space of the room. This helps us understand that the content analysis of visual data is useful to understand the visitors' experiences, while taking into account the specificities of installations. This is relevant, given that –as Budge [8] has argued, often times, the analysis of visitors' experiences is at the expense of understanding the object of installations.

For the artists and designers behind *Default*, analyzing the visitors' experiences through the visual content they shared on Instagram proved to be enlightening. For example, the analysis of the framing of the posts made visible the occurrence of a virtual division of the space that led to configuring different roles among the visitors: audience and performer. In retrospect, this might seem obvious; however, it was not clear for the developer team during the design process and offers an example of how analyzing the users' experiences can lead to modifications in further iterations. If *Default* was to be shown again, projections should be up all the walls of the room, to avoid such virtual division and to enhance even more the immersive experience. Thus, this experience not only offers evidence that the analysis of visual content shared on Instagram by visitors can be a useful tool for other designer/artists, but also, gives an example of how analyzing visitors' experiences opens avenues to further include intended audiences in the designing process. This, in turn, could help to effectively achieve the goals of sci-art, in terms of building bridges between the scientific community and non-expert audiences, and to democratize the designing process, helping artistic installations to better serve a social function [5, 25].

Among the limitations of this study we can count that Instagram users tend to be younger than 35 years old and therefore, it is not a platform appropriated to access nor analyze the experiences of older audiences. Thus, we cannot generalize the findings beyond this age group. Another limitation is the relatively small number of analyzed posts (53). A greater number could help us develop more conclusive findings, particularly given the possibility to access them easily. Given that the team did not provide in advance a hashtag, geotag or other identifier for the post related to the installation, it is possible to think that we missed posts.

The above leads to recommend that future works that might use Instagram as a platform to access audiences' experiences, is to provide both a hashtag and a geotag to unify the way of referencing the installation and easily access related Instagram posts. If various hashtags were provided (related with different aspects of the installation –for example: #defaultatMAC; #defaultSantiago; #defaultmodenetwork; #defaultinteractive; #defaultsciart, among others possible) the analysis of the hashtags chosen and caption written by the users could be utilized to further understand the aspects of the experience that they were emphasizing.

The process of analysis we propose in this work could be synthesized as follows: (1) Provide a hashtag at the exhibition; (2) Search for the hashtag provided; (3) Identify all the Instagram posts and then download all the pictures and videos; (4)Analyze the pictures and then videos, separately. We recommend that this step and the following be performed by at least two researchers in order to increase the rigor and reliability of the process; (5) Analyze pictures together, with focus on the content to identify recurring patterns, themes and differences; (6) Analyze videos together, with focus on the content to identify recurring patterns, themes and differences and (7) Analyze pictures and videos together to capture proto-themes that could turn into higher-order categories, to give account for the overall visitors' experiences.

In summary, we conclude that the visual analysis of Instagram posts shared is an effective way to identify and understand the impact of humanities, arts and culture in the web and also to develop the understanding about social media and its uses by web users and communities.

Finally, considering the growth and ease of accessibility of tools available for image processing, which can automatically perform image segmentation (e.g. SegNet [3]), there is room to increase the scale of this type of analysis for larger audiences.

7 ACKNOWLEDGMENTS

Default was funded by the Dirección de Artes y Cultura of the Vicerrectoría de Investigación at Pontificia Universidad Católica de Chile through the Arts & Technology, Science and Humanities 2016 contest and also supported by FONDECYT project N° 1140996, Anillo project ACT 1414, the CONICYT doctoral studies scholarship and the UC School of Design.

REFERENCES

[1] Ikuko, Acosta. (2001). Rediscovering properties inherent in art. *American Journal of Art Therapy*, 39 , 1 (Feb.2001) 93-97.

[2] M. Alejandro, alejandromerino.l. (2017).✨✨. Retrieved December 17, 2017 from https://www.instagram.com/p/BabXYlLnrNf/?taken-by=bryanmerino.l

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481-2495.

[4] S Barrass, and A. Sanchez (2014). *Experience and Evaluation in the Collective Creation of a Public Digital Exhibition.* In L. Candy and S. Ferguson, ed., Interactive experience in the digital age: evaluating new art practice. 209-219. Springer, Sydney.

[5] A. de Botton, and J. Armstrong (2013). *Art as therapy*. Phaidon, New York.

[6] Z. Bilda, L. Candy, and E. Edmonds (2007). *An embodied cognition framework for interactive experience.*. CoDesign 3, 2, 123-137

[7] Zafer Bilda, Chris Bowman, and Ernest Edmonds. 2008. Experience evaluation of interactive art: study of GEO landscapes. In *Proceedings of the 5th Australasian Conference on Interactive Entertainment* (IE '08). ACM, New York, NY, USA, Article 1, 10 pages. DOI: https://doi.org/10.1145/1514402.1514403

[8] G. Born and A. Barry. (2010). ART-SCIENCE. *Journal of Cultural Economy* 3,1, 103-119.

[9] K. Budge (2017). Objects in focus: Museum visitors and instagram. Curator: *The Museum Journal*, 60, 1, 67-85.

[10] L. Candy and S. Ferguson (2014). *Interactive experience in the digital age: evaluating new art practice*. Springer, Sydney.

[11] L. Candy. and E. Edmonds (2011). *Interacting: art, research and the creative practitioner*. Libri.

[12] A. Coles and G. Knowles, 2008. Arts-informed research. In G. Knowles & A. Cole (Eds). *Handbook of the arts in qualitative research.* Sage Publications, Los Angeles, CA, United States

[13] Cristian, cris.ignaciocm. (2017). Deja de pensar y abre tus sentidos. Retrieved December 17, 2017 from https://www.instagram.com/p/BZ4eygVlc9S/?taken-by=crissnacio

[14] C. Daniela, L. dacontrerasl. (2017). Default, de la diseñadora Manuela Garretón y el neurocientífico Tomás Ossandón. Retrieved December 17, 2017 from https://www.instagram.com/p/BZehf00ntoR/?taken-at=1314752441907553

[15] R. Diego, diegorubio83. (2017). Divagando - Mind Wandering.

[16] E. Edmonds (2014). Chapter 2: Human Computer Interaction, Art and Experience. In L. Candy and S. Ferguson, ed., *Interactive experience in the digital age: evaluating new art practice*. Springer, Sydney.

[17] M, Garretón, T. Ossandón and K. Hyland, 2018. Default. Design and Neuroscience: Interactive Work on Brain Activity. Diseña, 12, 286-299

[18] M. Garretón, K. Hyland. and D. Parra 2017. Understanding People's Interaction with Neural Sci-Art. *Proceedings of the IEEE VIS Conference, Arts Program VISAP '17.*

[19] R. Gingrich, A. Emet, E and Z. Xiao, 2014. "Transmission: A Telepresence Interface for Neural and Kinetic Interaction", *Leonardo*, 47, 4, 375-385. DOI https://doi.org/10.1162/LEON_a_00843

[20] K. Gleeson. (2011). Polytextual thematic analysis for visual data. Pinning down the analytic. In Reavey, P. (Ed.). *Visual methods in psychology. Using and interpreting images in qualitative research*. Psychology Press, New York.

[21] N. Hochman. and R. Schwartz, 2012. Visualizing Instagram: Tracing Cultural Visual Rhythms. AAAI Technical Report WS-12-03 Social Media Visualization.

[22] Instagram. (2018). Our Story. Retrieved January 14, 2019 from https://instagram-press.com/our-story/

[23] Rachel Jacobs, Steve Benford, Ewa Luger, and Candice Howarth. 2016. The Prediction Machine: Performing Scientific and Artistic Process. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16).* ACM, New York, NY, USA, 497-508. DOI: https://doi.org/10.1145/2901790.2901825

[24] Rachel Jacobs, Steve Benford, Mark Selby, Michael Golembewski, Dominic Price, and Gabriella Giannachi. 2013. A conversation between trees: what data feels like in the forest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* ACM, New York, NY, USA, 129-138. DOI: https://doi.org/10.1145/2470654.2470673

[25] E. Kramer. (2000). *Art as therapy. Collected papers*. Edited by Gerity, L. Jessica Kingsley Publishers. Philadelphia.

[26] L. Manovich (2017). Instagram and Contemporary Image. Retrieved January 14, 2019 from http://manovich.net/content/04-projects/147-instagram-and-contemporary-image/instagram_book_manovich.pdf

[27] C.H. Moon (2010) (Ed.). *Materials and Media in Art Therapy: Critical Understandings of Diverse Artistic Vocabularies*. Routledge. Florence, KY.

[28] Oda M, odamoura. (2017). Que confusão gostosa!bjus. Retrieved December 17, 2017 from https://www.instagram.com/p/Balws25AHBe/

[29] Pipexs_jt. (2017) Un breve timelapse de una presentación artística. Retrieved December 17, 2017 from https://www.instagram.com/p/BZ2BSXsj1vh/?taken-by=pipexs_jt

[30] M. Raichle, A. MacLeod, A. Snyder, W. Powers, D. Gusnard and G. Shulman, 2001. A Default mode of brain function. In *Proceedings of the National Academy of Sciences* 98, 2

[31] P. Ravey and K. Johnson, 2008. Visual approaches: using and interpreting images. In Willig, C. & Stainton-Rogers, W. (Eds). *The Sage handbook of Qualitative research in psychology*.Sage Publications Ltd. London.

[32] P. Reavey (2011). The return to experience. Psychology and the visual. In Reavey, P. (Ed.). *Visual methods in psychology. Using and interpreting images in qualitative research*.New York: Psychology Press.

[33] C.K. Riessman (2008) Visual analysis. In *Narrative methods for the human sciences*. Sage Publica, Los Angeles, 141-182.

[34] G. Rose (2002) Visual methodologies: *An introduction to the interpretation of visual materials*. Sage, London.

[35] Statista. (2018). Distribution of Instagram users worldwide as of January 2018, by age group. Retrieved January 14, 2019 from https://www.statista.com/statistics/325587/instagram-global-age-group/

[36] P. Toscano. (2017). Instagram-City: New Media, and the Social Perception of Public Spaces. *Visual Anthropology* 30, 3, 275-286.

[37] A. Weilenmann, T. Hillman and B. Jungselius, 2013. Instagram at the museum. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '13.

[38] S. Wilson (2002). Information Arts: Intersections of Art, Science, and Technology (Leonardo). MIT Press.

[39] R.West, J. Burke, C. Kerfeld, E. Mendelowitz, T. Holton, J. Lewis, E. Drucker and W. Yan. (2005). "Both and Neither: in silico v1.0, Ecce Homology", *Leonardo*, 38, 4, 286-293

Global Reactions to the Cambridge Analytica Scandal: A Cross-Language Social Media Study

Felipe González
UTFSM
Santiago, Chile
felipe.gonzalezpi@usm.cl

Yihan Yu
University of Washington
Seattle, United States
yyu2016@uw.edu

Andrea Figueroa
UTFSM
Valparaíso, Chile
andrea.figueroa@usm.cl

Claudia López
UTFSM
Valparaíso, Chile
claudia@inf.utfsm.cl

Cecilia Aragon
University of Washington
Seattle, United States
aragon@uw.edu

ABSTRACT

Currently, there is a limited understanding of how data privacy concerns vary across the world. The Cambridge Analytica scandal triggered a wide-ranging discussion on social media about user data collection and use practices. We conducted a cross-language study of this online conversation to compare how people speaking different languages react to data privacy breaches. We collected tweets about the scandal written in Spanish and English between April and July 2018. We used the Meaning Extraction Method in both datasets to identify their main topics. They reveal a similar emphasis on Zuckerberg's hearing in the US Congress and the scandal's impact on political issues. However, our analysis also shows that while English speakers tend to attribute responsibilities to companies, Spanish speakers are more likely to connect them to people. These findings show the potential of cross-language comparisons of social media data to deepen the understanding of cultural differences in data privacy perspectives.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Information systems** → *Document topic models.*

KEYWORDS

Data privacy; cross-language comparisons; Twitter

ACM Reference Format:
Felipe González, Yihan Yu, Andrea Figueroa, Claudia López, and Cecilia Aragon. 2019. Global Reactions to the Cambridge Analytica Scandal: A Cross-Language Social Media Study. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316456

1 INTRODUCTION

While there is evidence that concerns about privacy and its intricate relationship with users' decisions to use social media have been rising among Americans [30], much less is known about these perspectives across the world. So far, most of the literature about privacy concerns has been focused on the United States and even though recently research has begun to be carried out examining European privacy perspectives after the General Data Protection Regulation (GDPR) implementation, privacy research in international settings is still needed [29]. A few survey studies of people from different pairs of countries have been conducted to address. This, and the results so farindicate that data privacy perspectives vary significantly across countries. Belgians self-reported lower levels of concerns about sensitive information leakage than people from the United States. Harris et al. attributed this difference to the privacy laws of the respondents' countries of origin [13, 28]. A survey of North American and Turkish freshmen living in similar residence hall settings [21] showed that Americans wished for more privacy in their hall rooms than Turkish students. In the context of e-commerce, another survey found that Italians tend to exhibit lower privacy concerns than Americans [10].

Unfortunately, relying solely on survey data for cross-cultural studies of data privacy has various limitations. Most of them focus only on two geographic regions and have limited sample size [33]. Additionally, most privacy surveys are only available in English and only a few of them have been translated to other languages [33]. Surveys of a multinational or global nature that can mitigate these limitations would be very costly, which makes it difficult to compare privacy attitudes more broadly.

As the Cambridge Analytica scandal unfolded and people became aware that the personal data of 87 million Facebook users were exposed without their consent and used by Cambridge Analytica to support political campaigns [22], thousands of people in different parts of the world expressed on social media their reactions to and reflections on the scandal, its relationship to data privacy, and its broader implications. Indeed, a movement to delete Facebook accounts emerged and the #deleteFacebook hashtag was trending for several days [25].

In this paper, we report on a study that observes Twitter activity about the Cambridge Analytica scandal in Spanish and English and proposes a methodology for cross-language comparison of social media text. We believe that our approach offers an alternative or complementary method to conduct studies on data privacy perspectives across speakers of different languages and may provide a roadmap for future cross-cultural research. As Twitter allows people

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, SFO, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316456

to express themselves freely and spontaneously and in different languages, it enables a unique opportunity to analyze multi-language large-scale data. These characteristics allow researchers to address some of the limitations of the survey-based methods described above, such as those related to language and sample size [36]. Our premise is that written communication can be a "window into culture and an external reflection of cultural values" [7]; therefore, what people write about the scandal in their own languages can reveal differences and similarities in data privacy concerns across users worldwide.

We summarize prior work on cross-language comparisons of social media in Section 2. Section 3 introduces our research question, Section 4 details our research method, Section 5 reports on our findings, and Section 6 offers a discussion of our results, its limitations, and future work. Finally, Section 7 provides our conclusions.

2 CROSS-LANGUAGE COMPARISONS OF USER-GENERATED WRITTEN CONTENT

Prior research has used social media text in different languages to make comparisons among people who speak these languages. An analysis of more than 62 million tweets compared the top 10 most common languages regarding the use of features such as: URLs, hashtags, mentions, replies and re-tweets [17]. The findings show that German-speaking users tend to include more URLs and hashtags in their tweets than other users, while Korean-speaking users are prone to reply to each other more often than speakers of other languages. Hong et al. argue that users of different languages use Twitter for different purposes. The German community often uses this platform for information sharing, while the Korean community employs it for conversational purposes. Another study analyzes tweets written by Americans in English and Japanese in their official language [1]. As opposed to Americans, Japanese people tweet more self-related messages and more messages about TV programs. In turn, Americans tweet more about their peers, sports and news.

Previous work has also explored how user-generated content can reveal different views of the same issues among people who write in different languages. The Meaning Extraction Method (MEM) was applied to compare posts from depression-related forums in Spanish and English [26]. MEM is used to discover the main topics in a corpus. A comparison of the resulting topics shows that English posts tend to use words that are more concrete and descriptive and the main topics are related to medicinal questions and concerns. Spanish posts use relatively more emotional words and the main topics are associated with sharing and disclosing information about relational concerns. Hecht and Gergle used the Explicit Semantic Analysis (ESA) algorithm to analyze pairs of terms from ten different Wikipedia language editions [14]. ESA indicates a score of semantic relatedness between two concepts. The findings reveal that "even when two language editions cover the same concept, they may describe that concept differently" [14]. For example, consider the pair "Germany" / "Saxony-Anhalt" (a state of Germany). In most languages this pair receives a high ESA score, but the algorithm detects no relation at all in Italian and Danish. This occurs because there are no articles that mention "Germany" and "Saxony-Anhalt" together in these two languages. Analyses of semantic networks and the salience of semantic concepts in articles about China in the Chinese and English versions of Wikipedia found dissimilarities in the semantic content of these two versions [19]. Articles in the Chinese version are framed from the perspectives of respecting authority, emphasizing harmony and patriotism. Articles in English are written from the perspective of Western-societies' core value of democracy. The English version contains critical attitudes toward the authority of the Chinese government and the Communist Party in terms of human rights and territorial dispute. According to Jiang et al. cultures, values, interests, situations and emotions of different language groups can explain these dissimilarities. The latter studies provide evidence in support of what is known as the Sapir-Whorf hypothesis, which indicates that the structure of anyone's native language influences the world-views she or he will acquire as she or he learns the language [20]. Thus, speakers of different languages could think, perceive reality and organize the world around them in different ways [18]. Inspired by this hypothesis, we seek to study whether people who speak a different language hold different views of a data privacy scandal, such as the Cambridge Analytica case.

3 RESEARCH QUESTION

Given the relative lack of data privacy research in international settings [29], our project aims to investigate the potential of social media text written in different languages as a source to compare data privacy views worldwide. Beyond the Sapir-Whorf hypothesis (explained in Section 2), prior privacy research has argued that language and country of residence can relate to diverse perspectives on privacy. Smith et al. [29] noted that "many languages, including those in European countries (e.g Russian, French, Italian), do not have a word for privacy and have adopted the English word". Belanger and Crossler argued that "individuals from different countries can be expected to have different cultures, values and laws, which may result in differences in their perceptions of information privacy and its impacts" [3].

As the Cambridge Analytica scandal sparked worldwide conversations (in diverse languages) on Twitter about this particular misuse of user data, these online public communications are useful sources to contrast views on the scandal itself, its relation to data privacy, and its implications. To start addressing our ultimate research goal, this paper focuses on the following research question: "Which are the shared and unique topics that emerge from the Twitter activity in Spanish and English about the Cambridge Analytica data misuse scandal?"

4 DATA AND METHODS

To answer our research question, we used Tweepy[1], a Python library for accessing to the standard realtime streaming Twitter API. Using this library we were able to capture tweets that include hashtags or keywords related to the Cambridge Analytica scandal or data privacy, such as: "#CambridgeAnalytica", "#DeleteFacebook", "Zuckerberg" and "Facebook privacy". The standard realtime streaming Twitter API returns a random sample of all public tweets that match the search keywords. We collected tweets written in Spanish and English between April 1st and July 10th, 2018. Overall, we collected more than 7.4 million tweets written in English and more than 470,000 tweets in Spanish (see Table 1). The English tweets

[1] http://www.tweepy.org/

were generated by about 1.8 million unique Twitter accounts while the Spanish tweets were produced by approximately 220,000 users. The difference between the number of tweets and users collected in English and Spanish suggests that English-speaking Twitter users tweeted more about this scandal using the selected keywords than Spanish-speaking users, although this may be explained by the greater volume of English tweets overall[2].

We cleaned our dataset in two ways. First, we removed all retweets to focus our study on original user opinions and avoid analyzing duplicates. This step downsized both datasets to approximately 20% of their original sizes. Second, we attempted to eliminate tweets generated by automated accounts so our study could indeed reflect people's opinions. Unfortunately, there is not yet an infallible mechanism to detect bots' activity on Twitter. We chose to use Botometer to identify potential bots [9]. Botometer implements a machine learning algorithm that has achieved high accuracy (0.94) in detecting both simple and sophisticated bots in prior work [34]. The algorithm has been trained to detect bots by analyzing Twitter accounts' metadata, their contacts' metadata, tweets' content and sentiment, network patterns, and activity time series. The result is a score that is based on how likely it is to be a bot. The score ranges from 0 to 1, where lower scores indicate that the account behaves like a human and higher scores signal bot-like behavior. Unfortunately, there is not yet agreement on a threshold that can reliably distinguish bots from humans. To define thresholds for our two datasets, we used the Ckmeans [35] algorithm to cluster the Botometer scores in each dataset into five groups, with the first cluster including the accounts with the lowest Botometer scores (more human-like) and the fifth group the users with the highest scores (more bot-like). We reasoned that the fourth and fifth clusters in each dataset were least likely to contain humans; therefore, we excluded them from our analysis. Given that the Botometer analysis is time-consuming, we analyzed only a sample of users. In this study, we focused on the users who contributed the highest number of tweets in our datasets. In the future we plan to analyze users who contribute less. Thus, we were able to classify 19,478 accounts in the Spanish dataset (40.6%) and 74,021 (12.9%) accounts in the English dataset. Accounts with a Botometer score higher than 0.4745 were labeled as bots in the Spanish dataset. Those with a score higher than 0.4849 were considered bots in the English dataset. These users and their tweets were removed from our analysis. As a result, our final Spanish dataset includes 15,531 users who tweeted 50,559 times about the Cambridge Analytica scandal. The English dataset comprises 60,491 accounts that generated 446,462 tweets about it. Table 1 details these figures.

Table 1: Size of the Spanish and English datasets before and after data cleaning

Dataset	Spanish		English	
	#Tweets	#Users	#Tweets	#Users
Total	472,363	222,352	7,476,988	1,846,542
Without retweets	106,656	47,951	1,572,371	574,452
Most active	70,393	19,478	741,694	74,021
Humans	50,559	15,531	446,462	60,491

[2]https://www.statista.com/statistics/267129/most-used-languages-on-twitter/

To identify key topics in the resulting Spanish and English datasets, we used the Meaning Extraction Method (MEM) [8]. MEM is a topic modeling technique that can infer "what words are being used together, essentially resulting in a dictionary of word-to-category mappings from a collection of texts"[5]. After applying principal component analysis over this dictionary it is possible to identify words that can be grouped into themes or topics. This method has been identified as well-suited for cross-cultural and cross-language research [5]. MEM has been used to find themes in different contexts, such as: mental health [26, 38]; personality [6, 12] and values [37].

We employed the Meaning Extraction Helper (MEH) software [4], tool that can automate the majority of the MEM process [5]. The software contains a default list of Spanish and English stopwords; all these words were removed. Also, the tool allowed us to apply text segmentation by whitespace, conduct lemmatization and run Twitter-aware tokenization. To assist in lemmatization tasks, a conversion list was used to fix common misspellings (e.g "hieght" to "height") and convert "textisms" (e.g, "bf" to "boyfriend"). No stemming algorithm was used. We computed the frequency of each unigram as the percentage of tweets that contain it. The 300 most frequent unigrams were kept. We obtained a csv file with values of 1 and 0 indicating the corresponding unigrams' presence or absence, respectively, for each tweet.

Principal component factor analysis (PCA) was run over the MEH results. PCA was performed with varimax rotation to ensure that all resulting components are independent from each other. We conducted PCA with 5, 8, 11, 30 and 100 components. In both the Spanish and English datasets, 11 components gave the best results, with fit based upon diagonal scores of 0.55 and 0.9, respectively. This metric is a goodness of fit statistic, where values closer to 1 indicate better fit. The selected components accounted for 9% of the total variance of the Spanish data and 13% of the total variance of the English data.

To obtain the most representative words of each resulting component, we selected the words with factor loadings above 0.1, as recommended in [5]. The words were sorted according to their contribution to the component. Additionally, a python script was used to identify the top-30 tweets most related to each component. Using the most representative words and tweets by component, two authors examined and conceptualized the theme represented by each component, assigned a representative name and determined its relevance to our research question.

Finally, we used GeoNames API[3] to geo-locate all human-like accounts (see Table 1). Table 2 reports the proportion of tweets and users by the top-10 countries in each of our datasets. Most accounts could not be geo-located. The remaining accounts reveal that Spain and the US account for the majority of tweets and users in the Spanish and English datasets, respectively.

5 RESULTS

The words that clustered together to form coherent themes in the English and Spanish corpora are available online[4]. Tables 3 and 4 report the seven words with the highest loadings by component

[3]http://www.geonames.org/

[4]https://github.com/gonzalezf/LA-WEB-Paper

Table 2: Top ten most frequent user location in the English and Spanish datasets

Spanish			English		
Country	% tweets	% users	Country	% tweets	% users
not found	43.5%	46.05%	not found	41.6%	44.9%
Spain	17.5%	16.43%	U.S	33.6%	31.1%
Mexico	10.3%	9.90%	U.K	6.2%	6.9%
Venezuela	5.6%	3.56%	India	3.3%	3.0%
Argentina	5.4%	5.41%	Canada	2.4%	2.3%
Colombia	2.8%	2.98%	Australia	1.1%	1.4%
U.S	2.3%	2.40%	France	1.0%	0.6%
Chile	2.0%	2.35%	Germany	0.9%	0.7%
Peru	1.3%	1.41%	U.A.E	0.6%	0.3%
Ecuador	1.2%	1.25%	Netherlands	0.5%	0.4%
Brazil	0.9%	0.34%	Ireland	0.4%	0.4%

in the English and Spanish datasets, respectively. The tables also show the proportion explained (PE) by each of them according to the factor analysis. This number is proportional to the number of tweets associated with each theme. Table 5 presents the key themes in Twitter activity about the Cambridge Analytica scandal in Spanish and English. The themes are ordered according to their relevance to our research question.

Three key themes emerge in both languages. Spanish and English speakers talk about:

- "Cambridge Analytica's impact on political issues"
- "Mark Zuckerberg's Senate hearing in the USA," and
- "General Data Protection Regulation."

However, differences appear in how these themes are articulated. In regard to the first topic, the scandal's connection to Russia was much less relevant for Spanish speakers than for English speakers. English tweets focus on how Russia might have used Cambridge Analytica to intervene in the 2016 US elections and the UK Brexit campaign. For example, this component includes the following tweet: *"@ianbremmer @billmaher @RealTimers You want to know how Brexit happened and Trump got to win? Cambridge Analytica, Bannon, Mercers and the Bot Farms in Russia. They started testing MAGA, Build the Wall, Lock Her Up, Anti-Muslim sentiment, Anti-Immigrant propaganda. Putin worked hard at it since 2012"*. On the other hand, the token *Russia* does not appear as a representative word of this theme in the Spanish dataset. Instead, Spanish tweets are centered on Cambridge Analytica's closure as a result of the scandal. This behavior could be explained by the users' country of residence and its closeness to a salient political issue related to Cambridge Analytica. The US accounts for the largest share of users who report their location in our English dataset (see Table 2). It is well known that many people from this country have apprehensions about Russia since the Cold War. Furthermore, recent research has found evidence that Americans express "continued mistrust of Russia and a majority think Russia tried to interfere in the 2016 election" [27]. We believe that this is a plausible reason to explain this difference between the English and Spanish tweets.

While both datasets include a topic about "Mark Zuckerberg's Senate hearing in the US", their tweets' verb tenses differ. English-speaking users tend to tweet about this topic in future or present tense. These tweets reflect either certain level of anticipation of the event or live reports on how the event was unfolding. An example of these tweets is: *"Facebook CEO Mark Zuckerberg will testify in front of a joint hearing of the Senate Judiciary and Commerce Committees today. Senators will demand answers from Zuckerberg about Facebooks failure to protect up to 87 million users' private information https://t.co/9qpdqxjM7F https://t.co/aLlZtMMVCC"*. On the other hand, Spanish-speaking users are more likely to discuss this issue in past tense. They comment on sentences that Zuckerberg said during the hearing, putting special emphasis on the moment when he assumed responsibility for what has happened. For instance, translated Spanish tweets state: *"From #Whoknows to #ThroughMyFault, the change of attitude of #MarkZuckerberg. 'It was my mistake, and I'm sorry, I started #Facebook and I'm responsible for what happens here': Mark Zuckerberg before the Congress of #EEUU https://t.co/UX2QwT7pFw"*, and *"Zuckerberg takes full blame for the abuse of Cambridge Analytica before the US Senate: 'It was my mistake, and I'm sorry' https://t.co/VjH9ocdLcB"*. The difference in tenses (English future/present and Spanish past) may be explained by a delay in news reporting (volume of the Spanish tweets in our dataset relating to this particular topic tended to peak about 24 hours after the peak occurred in the English tweets) and translation to another language as the event occurred in an English-speaking country.

In the case of the "General Data Protection Regulation" (GDPR) theme, a tendency to attribute responsibility to companies for not attempting to comply with the GDPR was present in the English data, but not in the Spanish tweets. The English dataset includes accusatory tweets to companies such as Facebook and Google for trying to dodge GDPR rules, e.g. *"Facebook and Google are pushing users to share private information by offering invasive and limited default options despite new EU data protection laws aimed at giving users more control and choice https://t.co/x0srYwZeUz"*, *"If a new European personal data regulation (aka #GDPR) went into effect tomorrow, almost 1.9bn #Facebook users around the world would be protected by it. The online social network is making changes that ensure the number will be much smaller.https://t.co/UXonA0mTCs #privacy https://t.co/gqFaMYq9Su"*, and *"Facebook moves billions of international user accounts to California to avoid European privacy law https://t.co/nC2ccf2ClR"*. On the other hand, Spanish-speaking users tweet about GDPR to describe it and inform local companies how to prepare for it. Some translated tweets belonging to this component are: *" General #Data Protection Regulation (#GPDR) is a new #law of the European Union and it will enter into force on May 25th. Do you know what it is? Do you know its characteristics? Is your #company ready? Get up to date by clicking on the following link! https://t.co/AeKPR4w6Q9 https://t.co/rdlzbrAU1n"*, and *"On May 25th, the General Protection #Data Regulation (GDPR) of the EU, one of the most modern regulations regarding personal data use by companies and institutions, begins to be applied. There will be some repercussions in #Chile. https://t.co/XNbmALUcdD"*

Three other emerging themes are unique to Spanish speakers and are relevant to our research question as they all refer to the founder of Facebook and its role in different aspects of the scandal. These topics are:

Table 3: Terms by themes in the Spanish dataset

#ID	PE (%)	Words with the highest loadings by component						
		word1	word2	word3	word4	word5	word6	word7
S1	13%	animals	lottery	predict	live	program	listen	win
S2	11%	cambridge	analytica	million	user	scandal	data	affect
S3	10%	congress	error	zuckerberg	sorry	mark	usa	senate
S4	9%	rgpd	protection	gdpr	data	regulation	privacy	dataprotection
S5	9%	marketing	digitalmarketing	analytic	publicity	digital	google	youtube
S6	9%	press	like	followers	platform	follow-us	come	digital
S7	9%	red	social	twitter	instagram	youtube	facebook	follow-us
S8	8%	parliament	european	zuckerberg	mark	ask	hearing	sorry
S9	8%	stop	ask	join	people	red	create	senate
S10	8%	communitymanager	socialmedia	blog	socialnetworks	news	work	facebook
S11	7%	message	messenger	user	send	million	remove	delete

Table 4: Terms by themes in the English dataset

#ID	PE (%)	Words with the highest loadings by component						
		word1	word2	word3	word4	word5	word6	word7
E1	24%	newyorkcity	newyork	nyc	ny	career	code	hire
E2	12%	rsi	btc	signal	min	eth	bitcoin	crypto
E3	10%	machinelearn	deeplearn	artificialintelligence	ml	robotic	dl	ai
E4	10%	chatbot	infosec	databreach	hack	cybersecurity	crypto	blockchain
E5	9%	iot	iiot	smartcity	digitaltransformation	innovation	infographic	startup
E6	7%	cambridge	analytica	election	campaign	voter	vote	brexit
E7	7%	data	user	facebook	privacy	access	personal	law
E8	6%	zuckerberg	mark	testify	ceo	congress	committee	senate
E9	6%	trump	president	donald	white	house	democrat	america
E10	5%	social	media	twitter	instagram	censorship	conservative	facebook
E11	5%	late	daily	thanks	bigdata	remove	social	ai

Table 5: Themes in the Spanish and English datasets

Spanish		English	
Theme	#ID	Theme	#ID
Cambridge Analytica's impact on political issues	S2	Cambridge Analytica's impact on political issues	E6
Mark Zuckerberg's Senate hearing in the US	S3	Mark Zuckerberg's Senate hearing in the US	E8
General Data Protection Regulation	S4	General Data Protection Regulation	E7
Zuckerberg in front of the European Parliament	S8	Facts and opinions about Donald Trump	E9
Opinions about Mark Zuckerberg	S9	Stop censorship on social media	E10
Facebook deletes Zuckerberg's private messages	S11	Cryptocurrencies	E2
Digital marketing	S5	Artificial intelligence	E3
Promoting subscription to social platforms	S7	Blockchain	E4
Promoting likes in social platforms	S6	Internet of things	E5
Lottery results	S1	News about privacy on social media	E11
News and random facts	S10	Hiring tech jobs in New York	E1

- "Zuckerberg in front of the European Parliament"
- "Opinions about Mark Zuckerberg," and
- "Facebook deletes Mark Zuckerberg's private messages."

The first topic focused on Zuckerberg's laments for the situation, e.g. *"Mark Zuckerberg apologized to the European Parliament for the data breach. The founder of Facebook acknowledged on Tuesday that the tools of the social network were used 'to do harm'. https://t.co/aeDUVViUhq"* We note that this theme only emerges in the Spanish tweets. Again, we attribute this distinction to the users' country of residence. Both datasets include users whose self-reported location is in Europe; however, they are the majority only in the Spanish dataset where Spain is associated with more tweets

than any other country (see Table 2). Thus, the hearing in the EU Parliament is much more prominent in the Spanish dataset.

The second topic contains supporting and accusatory tweets for Mark Zuckerberg in relation to the Cambridge Analytica scandal. This theme includes tweets such as the following: *"what they do not forgive to Zuckerberg is that Facebook has unwittingly helped Trump triumph (which I do not think), because when Facebook censored pages of right-wing groups for nothing (the same Mark Zuckerberg has said that he worries about the leftist prejudice of his staff), there they said nothing."*

The third theme groups together tweets about the option to delete private direct messages on Facebook. This functionality apparently was available to Mark Zuckerberg in the past while no other user could use it. Spanish speakers tweet about the special privileges of Zuckerberg as a Facebook user and about the possibility that such functionality could become available to everyone. Example translated tweets are: *"Although you can not delete your messages sent from another person's inbox in Messenger, Facebook has done so in the case of Mark Zuckerberg and other company executives. https://t.co/PkJByaesdT"*, and *"It is clear that Mark Zuckerberg is the only one who can have the God Mode of Facebook because only he can erase messages sent through Facebook Messenger, a feature that will be released for everyone after this scandal."*

Two themes are unique to English and are considered slightly relevant to our research question. They are conceptualized as:

- "Stop censorship on social media," and
- "News about privacy on social media."

First, English-speaking users protest about censorship on social media companies such as Facebook, Twitter, Instagram and Google. They demand freedom of speech. This component includes the following tweet: *"@facebook Social media censorship is tearing out our tongues! Is Facebook a neutral public forum? STOP CENSORSHIP! Demand an #InternetBillOfRights #IBOR #MAGA #InvasionOfPrivacy #HumanRights #DeleteFacebook #Censorship #Twitter #Censorship #Google #FaceBook #Instagram @realDonaldTrump "*. The next topic includes informative tweets (usually from electronic newspapers) that relate to social media and privacy, e.g. *"The latest The GDPR and Data Protection Daily! https://t.co/GZBscXkxPl Thanks to @simpledatainc @Colt_Technology @content_app #bigdata"*.

Other key themes arise from both datasets, but we consider them not sufficiently relevant to our research question. In the Spanish dataset, these themes are: "Digital marketing", "Promoting likes in social platforms", "Promoting subscription to social platforms", "Lottery results", and "News and random facts". In English, there are the following topics: "Cryptocurrencies", "Artificial Intelligence", "Blockchain", "Internet of Things", "Facts and opinions about Donald Trump," and "Hiring tech jobs in New York."

6 DISCUSSION

Our study of tweets in Spanish and English about the Cambridge Analytica data misuse scandal allowed us to conduct a cross-language comparison of their main emerging themes. Out of eleven topics, three are common to both languages. However, they present meaningful differences in their articulation. Three other relevant themes are unique to the Spanish data and two relevant topics only arise from the English tweets. These findings show the potential of our algorithm-based approach for cross-language comparisons of tweets to identify similarities and (nuanced) differences on privacy-related views across people from different countries.

We discuss here two underlying patterns that could explain the distinctions we found between Spanish and English speakers.

First, we observe a tendency for a local perspective to appear in tweets about this data privacy scandal. Tweets in English provide elaborate rationales behind the Cambridge Analytica's impact on political issues. They often relate them to Russia's actions. This is a common argument among Americans [27], who are also the most active contributors of tweets in our English dataset. This kind of rationale is much less visible in tweets in the Spanish dataset, where most users are not located in the US. Furthermore, there is a large component of tweets about Mark Zuckerberg's hearing in front of the US Senate in both languages; however, only the Spanish corpus has a theme related to Zuckerberg's audience in the European parliament. Again, we relate this finding to the users' country of residence. Our Spanish dataset consists primarily of European users. This could explain why this event became more salient only in this language. Another finding that can be explained by a local perspective is the demand for freedom of speech in the English tweets. This theme does not emerge at all in the Spanish data. Free speech is a central value in the US [2], but it is not as prominent in Spanish-speaking countries.

Second, English tweets often attribute responsibilities to governments and organizations while Spanish tweets tend to question individual actions. For example, English-speaking users tend to tweet accusatory statements about big companies such as Google and Facebook dodging GDPR. On the other hand, these large companies are very rarely questioned about their acts in the Spanish data. Instead, it is possible to find many tweets that support Mark Zuckerberg and blame Facebook users for being irresponsible by failing to read Facebook's terms of service and neglecting to protect their personal data.

Research on cross-cultural differences could provide a framework to further explore the second pattern. This research argues that national cultures could be characterized by their scores on a small number of dimensions [24]. Hofstede's cultural dimensions [15] have been widely used to study the relationship between people's culture and technology [24]. We hypothesize that one of these dimensions, "power distance", could explain differences between Spanish and English speakers on responsibility attribution. This dimension is defined as: "the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally" [16]. In countries with a high level of power distance, people tend to accept hierarchies without further justification and authority is hardly questioned. Most Anglo and Nordic countries are characterized by small power distance scores. The opposite was found in Latin American countries. These nations often score high in power distance [15]. This pattern could explain why many tweets that question authorities (e.g., governments and companies) occur in English while they are much less present in Spanish. While this is a plausible path of reasoning, cross-cultural research has its own limitations [31, 32]; therefore, additional work is needed to confirm or refute this explanation.

In both the English and the Spanish dataset we found tweets from spammers and self-promoters' accounts. These accounts tend

to use hashtags relevant to popular online discussions (e.g: #Facebook, #privacy) to promote their conversations and possibly share malicious links. These tweets account for the themes that were considered irrelevant in our analysis.

6.1 Limitations and future work

As in any study, our research has limitations that need to be taken into consideration. We collected data through the standard streaming Twitter API and by using specific hashtags and keywords. Thus, we only had access to a small sample of all the tweets about the scandal. Nevertheless, studies have shown that this API is highly correlated with the full Twitter stream [23]. While we attempted to identify and remove bot activity from our data, and the threshold we defined lies between the recommended range [0.43-0.49] [34], our dataset contains both false positives and negatives, which could affect the results. Also, difficulties interpreting MEM results could occur due to the lack of contextual information such as "information about valence and timing that is obscured in the analysis" [11]. Further qualitative analysis may help with future interpretation.

Future work includes investigating the English and Spanish corpora in order to strengthen the analysis of shared themes. This could be done, for example, by incorporating measurements of similarity between texts. Using bigrams and trigrams as inputs of the MEM analysis could help to contextualize the topic components. Also, it could be interesting to compare MEM results with other topic modeling techniques such as LDA or LDA2Vec. Furthermore, exploring the data over the time dimension should be considered because it may allow the detection of topic changes over time.

7 CONCLUSIONS

We proposed a methodology to conduct a cross-language study of Twitter activity about a major data privacy leakage: the Cambridge Analytica scandal. We collected tweets about it in English and Spanish during 100 days, processed them to remove tweets generated by bots, and conducted MEM to identify the key themes in each language. We conceptualized the themes and compared them. We identified five topics that were unique to only one of the languages. We also found three common topics: "Cambridge Analytica's impact on political issues," "Mark Zuckerberg's Senate hearing in the USA," and "General Data Protection Regulation." Nevertheless, we detected dissimilarities in how these topics were formulated in each language. We proposed two plausible patterns that may explain these differences. First, reactions to a data privacy scandal reflect the users' local perspectives. Second, there is a tendency for English speakers to assign responsibilities to governments and organizations while Spanish speakers tend to attribute them to people.

The contributions of our work and findings are twofold. First, we proposed a method to leverage social media data to investigate reactions to a data privacy scandal in two languages that cover a multi-national audience. Second, we found cross-language differences on privacy-related views that go beyond measuring levels of privacy concerns, as defined in widely-adopted surveys such as [30]. We expect that future work can deepen the understanding of these differences through diverse methods of research.

ACKNOWLEDGMENTS

This collaboration was possible thanks to the support of the Fulbright Program, under a 2017-18 Fulbright Fellowship award. This work was also partially funded by CONICYT Chile, under grant Conicyt-Fondecyt Iniciación 11161026. The first author acknowledges the support of the PIIC program from Universidad Técnica Federico Santa María and CONICYT-PFCHA/MagísterNacional/2019 - 22190332.

REFERENCES

[1] Adam Acar. 2013. Culture and social media usage: Analysis of Japanese Twitter users. *International Journal of Electronic Commerce Studies* 4, 1 (2013), 21–32.
[2] Mauricio J Alvarez and Markus Kemmelmeier. 2018. Free Speech as a Cultural Value in the United States. *Journal of Social and Political Psychology* 5, 2 (2018), 707–735.
[3] France Bélanger and Robert E Crossler. 2011. Privacy in the digital age: a review of information privacy research in information systems. *MIS quarterly* 35, 4 (2011), 1017–1042.
[4] RL Boyd. 2014. MEH: Meaning Extraction Helper (Version 1.0. 6)[Software].
[5] Ryan L Boyd. 2017. Psychological text analysis in the digital humanities. In *Data analytics in digital humanities*. Springer, 161–189.
[6] Ryan L Boyd and James W Pennebaker. 2015. Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological science* 26, 5 (2015), 570–582.
[7] Kyungsub Stephen Choi, Il Im, and Gert Jan Hofstede. 2016. A cross-cultural comparative analysis of small group collaboration using mobile twitter. *Computers in Human Behavior* 65 (2016), 308–318.
[8] Cindy K Chung and James W Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality* 42, 1 (2008), 96–132.
[9] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 273–274.
[10] Tamara Dinev, Massimo Bellotto, Paul Hart, Vincenzo Russo, Ilaria Serra, and Christian Colautti. 2006. Privacy calculus model in e-commerce–a study of Italy and the United States. *European Journal of Information Systems* 15, 4 (2006), 389–402.
[11] Marilyn R Fitzpatrick and Calli Renee Armstrong. 2010. Beyond the tip of the iceberg: Exploring the potential of the meaning extraction method and aftercare e-mail themes. *Psychotherapy research* 20, 1 (2010), 86–89.
[12] Eden M Griffin and Karen L Fingerman. 2018. Online Dating Profile Content of Older Adults Seeking Same-and Cross-Sex Relationships. *Journal of GLBT Family Studies* 14, 5 (2018), 446–466.
[13] Michael M Harris, Greet Van Hoye, and Filip Lievens. 2003. Privacy and attitudes towards internet-based selection systems: a cross-cultural comparison. *International Journal of Selection and Assessment* 11, 2-3 (2003), 230–236.
[14] Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 291–300.
[15] Geert Hofstede. 1983. National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization* 13, 1-2 (1983), 46–74.
[16] Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.
[17] Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language Matters In Twitter: A Large Scale Study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 518–521.
[18] Basel Al-Sheikh Hussein. 2012. The Sapir-Whorf Hypothesis Today. *Theory and Practice in Language Studies* 2, 3 (2012), 642–646.
[19] Ke Jiang, Grace A. Benefield, Junfei Yang, and George Barnett. 2017. Mapping Articles on China in Wikipedia An Inter-Language Semantic Network Analysis. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. IEEE.
[20] Paul Kay and Willett Kempton. 1984. What is the Sapir-Whorf hypothesis? *American anthropologist* 86, 1 (1984), 65–79.
[21] Naz Kaya and Margaret J Weber. 2003. Cross-cultural differences in the perception of crowding and privacy regulation: American and Turkish students. *Journal of environmental psychology* 23, 3 (2003), 301–309.
[22] Jean-Rémi Lapaire. 2018. Why content matters. Zuckerberg, Vox Media and the Cambridge Analytica data leak. *ANTARES: Letras e Humanidades* 10, 20 (2018), 88–110.
[23] Kalev Leetaru. 2019. Is Twitter's Spritzer Stream Really A Nearly Perfect 1% Sample Of Its Firehose?

[24] Dorothy E Leidner and Timothy Kayworth. 2006. A review of culture in information systems research: Toward a theory of information technology culture conflict. *MIS quarterly* 30, 2 (2006), 357–399.

[25] Maya Mirchandani. 2018. To delete or not to# Delete Facebook, that is the question.

[26] Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*.

[27] Dina Smeltz and Lily Wojtowicz. 2018. Public Opinion on US-Russian Relations in the Aftermath of the 2016 Election. *SAIS Review of International Affairs* 38, 1 (2018), 17–26.

[28] H Jeff Smith. 2001. Information privacy and marketing: What the US should (and shouldn't) learn from Europe. *California Management Review* 43, 2 (2001), 8–33.

[29] H Jeff Smith, Tamara Dinev, and Heng Xu. 2011. Information privacy research: an interdisciplinary review. *MIS quarterly* 35, 4 (2011), 989–1016.

[30] H Jeff Smith, Sandra J Milberg, and Sandra J Burke. 1996. Information privacy: measuring individuals' concerns about organizational practices. *MIS quarterly* (1996), 167–196.

[31] Ralf Terlutter, Sandra Diehl, and Barbara Mueller. 2006. The GLOBE study—applicability of a new typology of cultural dimensions for cross-cultural marketing and advertising research. In *International advertising and communication*. Springer, 419–438.

[32] Rosalie L Tung and Alain Verbeke. 2010. Beyond Hofstede and GLOBE: Improving the quality of cross-cultural research. *Journal of International Business Studies* 41, 8 (01 Oct 2010), 1259–1274.

[33] Blase Ur and Yang Wang. 2013. A cross-cultural framework for protecting user privacy in online social media. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 755–762.

[34] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).

[35] Haizhou Wang and Mingzhou Song. 2011. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R journal* 3, 2 (2011), 29.

[36] Yuanyuan Wang, Muhammad Syafiq Mohd Pozi, Yukiko Kawai, Adam Jatowt, and Toyokazu Akiyama. 2017. Exploring Cross-cultural Crowd Sentiments on Twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 321–322.

[37] Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 143–152.

[38] Markus Wolf, Cindy K Chung, and Hans Kordy. 2010. Inpatient treatment to online aftercare: e-mailing themes as a function of therapeutic outcomes. *Psychotherapy Research* 20, 1 (2010), 71–85.

Implementation of the Best Practices for Data on the Web in Brazil and Costa Rica

Bernadette Farias Lóscio
bfl@cin.ufpe.br
Center of Informatics - Federal University of Pernambuco
Recife, Brazil

Caroline Burle
cburle@nic.br
Brazilian Network Information Center - NIC.br
São Paulo, Brazil

Newton Calegari
newton@nic.br
Brazilian Network Information Center - NIC.br
São Paulo, Brazil

ABSTRACT

The Best Practices described on the Data on the Web Best Practices (DWBP) document [3] encourages and enables the continued expansion of the Web as a medium for the exchange of data. In this context, this paper focus on two cases of implementing the DWBP. The first one concerns data published by The Regional Center for Studies on the Development of the Information Society (Cetic.br) of The Brazilian Network Information Center (NIC.br). The second use case shows the experience of the Judiciary Department of Costa Rica (Justicia Abierta) with applying the DWBP Recommendation to publish their data on the Web.

KEYWORDS

Data on the Web, Best Practices, Application in Brazil and Costa Rica, Open Data

ACM Reference Format:
Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. 2019. Implementation of the Best Practices for Data on the Web in Brazil and Costa Rica. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316463

1 INTRODUCTION

One of the main goals of the Data on the Web Best Practices (DWBP) [3] is to facilitate interaction between publishers and consumers of data on the Web. A set of 35 Best Practices was created by W3C as a Web Recommendation to cover different challenges related to data publishing and consumption, such as Metadata, Data licenses, Data provenance, Data quality, Data versioning, Data identification, Data formats, Data vocabularies, Data access and APIs, Data preservation, Feedback, Data enrichment and Data republication.

In this context, this paper aims to illustrate the use of the DWBP through real use cases. For this, we present the Cetic.br [1] and the Justicia Abierta [2] use cases for applying the set of 35 Best Practices to publish their data on the Web. These two use cases illustrate some of the main challenges faced when applying the DWBP. On the one hand, their evaluation about the DWBP shows that a large number of best practices are easier to implement, on the other hand, both scenarios found difficulties in applying a subset of the best practices.

The remainder of this paper is organized as follows. Section 2 presents an overview of the DWBP. Section 3 presents the Brazil use case. Section 4 presents the Costa Rica use case. Section 5 presents some conclusions.

2 DATA ON THE WEB BEST PRATICES

The Best Practices described on the Data on the Web Best Practices document [3] encourages and enables the continued expansion of the Web as a medium for the exchange of data. In broad terms, data publishers aim to share data either openly or with controlled access. Data consumers (who may also be publishers themselves) want to be able to find, use and link to the data, especially if it is accurate, regularly updated and guaranteed to be available at all times. This creates a fundamental need for a common understanding between data publishers and data consumers. Without this agreement, data publishers' efforts may be incompatible with data consumers' desires.

In this context, it becomes crucial to provide guidance to publishers that will improve consistency in the way data is managed. Such guidance will promote the reuse of data and foster trust in the data among developers, whatever technology they choose to use, increasing the potential for genuine innovation.

The set of Best Practices proposed in [3] was developed to cover the lack of technical guidance for publishing data on the Web. As described in [3], each Best Practice has an intended outcome, which describes "*What it should be possible to do when a data publisher follows the best practice*". In general, the intended outcome is an improvement in the way that a data consumer (human or software) can manipulate a dataset published on the Web. In some cases, the intended outcome reflects an improvement in the dataset itself, which will also result in a gain for the data consumer.

In order to encourage data publishers to adopt the DWBP, the recommendation proposes a number of distinct DWBP benefits, including: comprehension; processability; discoverability; reuse; trust; linkability; access; and interoperability. The benefits are important to engage data publishers in using the Best Practices because they help them to have a better understanding of "what will be possible" when one or more Best Practices are adopted.

Each benefit is associated to one or more specific Best Practices. For example, the benefit "Comprehension" is associated to ten Best Practices, which are related to metadata, data vocabularies, feedback and data enrichment. This means that if a data publisher adopts these Best Practices, the level of comprehensibility will increase, i.e., it will be possible for humans to have a better understanding

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316463

about the data structure, the data meaning, the metadata and the nature of the dataset.

The following list shows the set of best practices linked to the DWBP document:

- Best Practice 1: Provide metadata
- Best Practice 2: Provide descriptive metadata
- Best Practice 3: Provide structural metadata
- Best Practice 4: Provide data license information
- Best Practice 5: Provide data provenance information
- Best Practice 6: Provide data quality information
- Best Practice 7: Provide a version indicator
- Best Practice 8: Provide version history
- Best Practice 9: Use persistent URIs as identifiers of datasets
- Best Practice 10: Use persistent URIs as identifiers within datasets
- Best Practice 11: Assign URIs to dataset versions and series
- Best Practice 12: Use machine-readable standardized data formats
- Best Practice 13: Use locale-neutral data representations
- Best Practice 14: Provide data in multiple formats
- Best Practice 15: Reuse vocabularies, preferably standardized ones
- Best Practice 16: Choose the right formalization level
- Best Practice 17: Provide bulk download
- Best Practice 18: Provide Subsets for Large Datasets
- Best Practice 19: Use content negotiation for serving data available in multiple formats
- Best Practice 20: Provide real-time access
- Best Practice 21: Provide data up to date
- Best Practice 22: Provide an explanation for data that is not available
- Best Practice 23: Make data available through an API
- Best Practice 24: Use Web Standards as the foundation of APIs
- Best Practice 25: Provide complete documentation for your API
- Best Practice 26: Avoid Breaking Changes to Your API
- Best Practice 27: Preserve identifiers
- Best Practice 28: Assess dataset coverage
- Best Practice 29: Gather feedback from data consumers
- Best Practice 30: Make feedback available
- Best Practice 31: Enrich data by generating new data
- Best Practice 32: Provide Complementary Presentations
- Best Practice 33: Provide Feedback to the Original Publisher
- Best Practice 34: Follow Licensing Terms
- Best Practice 35: Cite the Original Publication

3 BRAZIL'S USE CASE

The Center for Studies on the Development of the Information Society (Cetic.br) produces indicators and statistics on the use of information and communication technologies in Brazil since 2005. In 2017, it started the process of providing microdata databases of its sample surveys. Doing this, the Cetic.br allows more people to access and use their data, as well as different devices could recognize and read their data, thus broadening the possibilities of use of the databases published by the Center [1].

Table 1: Assessing the implementation of DWBP on Cetic.br microdata website

Dimension	Best Practice	Applies	Applies immediately
Metadata	BP01	Yes	Yes
	BP02	Yes	Yes
	BP03	Yes	Yes
Data Licenses	BP02	Yes	Yes
Data Provenance	BP05	Yes	Yes
Data Quality	BP06	No	No
Data Versioning	BP07	Yes	Yes
	BP08	Yes	Yes
Data Identifiers	BP09	Yes	Yes
	BP10	No	No
	BP11	No	No
Data Formats	BP12	Yes	Yes
	BP13	No	No
	BP14	Yes	Yes
Data Vocabularies	BP15	Yes	Yes
	BP16	No	No
Data Access	BP17	Yes	Yes
	BP18	No	No
	BP19	Yes	No
	BP20	No	No
	BP21	Yes	Yes
	BP22	Yes	Yes
Data Access APIs	BP23	No	No
	BP24	No	No
	BP25	No	No
	BP26	No	No
Data Preservation	BP27	Yes	Yes
	BP28	No	No
Feedback	BP29	Yes	Yes
	BP30	Yes	Yes
Data Enrichment	BP31	No	No
	BP32	Yes	Yes
Republication	BP33	No	No
	BP34	No	No
	BP35	No	No

In order to improve the quality of their data publication process, the Center for Studies on the Development of the Information Society (Cetic.br) followed the steps described below to apply the DWBP:

(1) Meetings with the teams of Cetic.br, Web Technology Study Center (Ceweb.br), Legal Department and Web Systems Department of NIC.br were held to level the knowledge of these teams on DWBP.
(2) A checklist[1] with the 35 Best Practices was made to understand which of them were applicable right away, which

[1]https://docs.google.com/spreadsheets/d/1yXyd3bnvLWL6KqKZzUbd8_BrB47YG8extgF3eBOHEDw/edit#gid=0

should be postponed and for how long, and the ones that were not applicable.
(3) The DWBP applicable were implemented on Cetic.br Microdata Website[2]

An important step in the process of applying the DWBP concerns the meetings with the teams of Cetic.br, Ceweb.br, Legal Department and Web Systems Department of of NIC.br to level the knowledge of these teams on DWBP. After the meetings, the necessary adaptations - both in the implementation of the BPs and in the processes of the research of Cetic.br -, were implemented. This was also a first step to apply the DWBP and to become part of Cetic.br's research process and data publication.

Table 1 shows the result of the checklist, meaning which BPs were applicable, which ones were not applicable and which ones should be adapted before being implemented.

The checklist presented in the Table 1 shows that 18 Best Practices could be applied immediately and the BP19 "Use content negotiation for serving data available in multiple formats" should be applied afterwards because the Web System Department of NIC.br needed more time to implement such BP.

Figure 1: Cetic.br microdata website

Most of the 16 BPs not applicable would not be implemented because of specificities of the research methodology and data production. For instance, BP20 "Provide real-time access" could not be implemented because of the data collection process, which does not allow to publish data in real time. Another example of BPs that could not be implemented are the ones in the Dimension of Republication: BP33 "Provide Feedback to the Original Publisher"; BP34 "Provide Feedback to the Original Publisher"; and BP35 "Cite the Original Publication". Regarding that Cetic.br is the data producer and does not republishes data from other statistical institutes.

According to [1], one of the lessons learned from the team involved in this process was that, despite being technical recommendations, it is not the technical character that limits its implementation. The main barrier is the engagement around publication and the perception of importance and benefits - for publishers and for data consumers - to implement the BPs and to publish and share the Center's databases. In this sense, the importance of involving people from different areas is highlighted, so as to promote an engagement and alignment between all the parties involved. The Figure 1 shows part of the Cetic.br microdata website front page.

4 COSTA RICA'S USE CASE

In 2017, the Judiciary Department of Costa Rica was supported by the The Trust for the Americas Foundation to create the Justicia Abierta (Open Justice) and open its data [2].

Similar to the Cetic.br, the Judiciary Department of Costa Rica followed the steps described below to apply the DWBP:

(1) The Web Technology Study Center (Ceweb.br) team ministered a course on publishing data on the web - based on the Data on the Web Best Practices (DWBP) - for twenty employees of the Government of Costa Rica, in February of 2017.
(2) The Open Justice Department created an interdisciplinary group to add knowledge on publishing data on the Web.
(3) The Justice Department team conducted a public consultation in partnership with the Abriendo Datos initiative to obtain from the Costa Rican society which data they should prioritize. They then collected, cleaned and checked those data.
(4) A checklist with the 35 Best Practices was made to understand which of them were applicable right away, which should be postponed and for how long, and the ones that were not applicable.
(5) The DWBP applicable were implemented on Justicia Abierta Website[3].

Table 2 shows the result of the checklist, meaning which BPs were applicable, which ones were not applicable and which ones would need to be adapted before implemented. The checklist shows that 16 Best Practices were to be applied immediately and 06 BPs should be applied afterwards. Technical limitations were the main reason for postponing the adoption of these 06 BPs.

The biggest challenge reported by the Justicia Abierta leader was the lack of knowledge on the part of public officials involved in the choice of those data and also some resistance by some guardians of the data. According to him, these difficulties were overcome by means of conversations and training that took place with the interdisciplinary group[4].

At the end, after the internal process have been established, of the 35 DWBP, the Justice Department of Costa Rica applied 17 BPs. It is noteworthy that, prior to step 1, only 3 BPs were adopted by the Costa Rica Justice Department.

[2]https://www.cetic.br/microdados/

[3]http://datosabiertospj.eastus.cloudapp.azure.com/

[4]http://ceweb.br/media/docs/publicacoes/10/DWBP_Condatos_Caroline%20Burle_Carlos%20Morales.pdf

Table 2: Assessing the implementation of DWBP on the Justice Department website

Dimension	Best Practice	Applies	Applies immediately
Metadata	BP01	Yes	Yes
	BP02	Yes	Yes
	BP03	Yes	Yes
Data Licenses	BP04	Yes	Yes
Data Provenance	BP05	Yes	Yes
Data Quality	BP06	Yes	Yes
Data Versioning	BP07	Yes	Yes
	BP08	Yes	Yes
Data Identifiers	BP09	Yes	No
	BP10	No	No
	BP11	Yes	No
Data Formats	BP12	Yes	Yes
	BP13	Yes	No
	BP14	Yes	Yes
Data Vocabularies	BP15	Yes	Yes
	BP16	Yes	No
Data Access	BP17	Yes	Yes
	BP18	Yes	No
	BP19	No	No
	BP20	No	No
	BP21	Yes	Yes
	BP22	No	No
Data Access APIs	BP23	No	No
	BP24	No	No
	BP25	No	No
	BP26	No	No
Data Preservation	BP27	Yes	No
	BP28	Yes	Yes
Feedback	BP29	Yes	Yes
	BP30	No	No
Data Enrichment	BP31	No	No
	BP32	Yes	Yes
Republication	BP33	No	No
	BP34	No	No
	BP35	No	No

The Figure 2 shows *Justicia Abierta* website open data front page.

5 DISCUSSION

This paper presented two real use cases of the DWBP application: the Cetic.br [1] and the Justicia Abierta [2] use cases. Both use cases adopted a similar methodology to implement the DWBP, including team meetings as well as a detailed evaluation of the DWBP.

Both Cetic.br and *Justicia Abierta* use cases demonstrated the importance of gathering all people involved in the process of publishing data. In the case of Cetic.br, even though the data was produced by the Center, holding meetings with the other departments involved such as the Legal Department and Web System Department of NIC.br, as well as with Ceweb.br to help the explanation about the DWBP, was very important to the process. In the case of

Figure 2: Costa Rica *Justicia Abierta* website

Open Justice, conversations and trainings that took place with the interdisciplinary group were significant to overcome the difficulties on lack of knowledge and resistance by some guardians of the data.

It is important to note that in both cases some Best Practices could be implemented immediately: 1 to 5 - about Metadata, Data Licenses and Data Provenance; 7 and 8 - about Data Versioning; 12 and 14 - about Data Formats; 15 - about Data Vocabularies; 17 and 21 - on Data Access; 29 - about Feedback; and 32 - about Data Enrichment. This shows that some of the proposed best practices can be easily adopted without major modifications on the data publication process. On the other hand, neither were able to implement the BPs 10 - about Data Identifiers; 20 - on Data Access; 23 to 26 - about Data Access APIs; 31 - about Data Enrichment; and 33 to 35 - about Republication. For implementing these BPs, more technical or advanced knowledge becomes necessary.

Although neither use cases implemented all the Data on the Web Best Practices, both use cases showed that it is possible to implement some of them without great effort. Furthermore, according to both teams, understand the DWBP as a guide and not as rules to be fulfilled is of great importance. We understand that both Brazil's and Costa Rica's use cases may be a reference to other data publishers of how to use the DWBP to facilitate interaction between data publishers and consumers.

REFERENCES

[1] Tatiana Jereissati Javiera Medina Macaya Beatriz Rossi Corrales, Caroline Burle. 2018. A importância das Boas Práticas para Dados na Web e o caso do Cetic.br. *GT.ADM09 Medición Y Evaluación En Apertura De Datos: Desafíos, Retos, Experiencias.* (2018).

[2] Caroline Burle. 2018. Implementação das boas práticas de dados na web: o caso da Costa Rica. (2018). [Online; Accessed 21 Jan 2019].

[3] Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. 2017. Data on the Web Best Practices. https://www.w3.org/TR/dwbp/. *W3C Recommendation* (2017). [Online; accessed 13 Jan 2019].

Mining the Relationship Between Car Theft and Places of Social Interest in Santiago Chile

Karen Oróstica
Department of Chemical Engineering, Biotechnology and Materials, University of Chile
Santiago, Chile
korostica@ing.uchile.cl

Barbara Poblete
IMFD - Department of Computer Science
University of Chile
Santiago, Chile
bpoblete@dcc.uchile.cl

ABSTRACT

Recent work suggests that certain places can be more attractive for car theft based how many people regularly visit them, as well as other factors. In this sense, we must also consider the city or district itself where vehicles are stolen. All cities have different cultural and socioeconomic characteristics that influence car theft patterns. In particular, the distribution of public services and places attract a large crowd could play a key role in the occurrence of car theft. Santiago, a city that displays drastic socioeconomic differences among its districts, presents increasingly-high car theft rates. This represents a serious issue for the city, as for any other major city, which –at least for Santiago– has not been analyzed in depth using quantitative approaches. In this work, we present a preliminary study of how places that create social interest, such as restaurants, bars, schools, and shopping malls, increase car theft frequency in Santiago. We also study if some types of places are more attractive than others for this type of crime. To evaluate this, we propose to analyze *car theft points* (CTP) from insurance companies and their relationship with *places of social interest* (PSI) extracted from Google Maps, using a proximity based approach. Our findings show a high correlation between CTP and PSI for all of the social interest categories that we studied in the different districts of the Santiago. In particular our work contributes to the understanding of the social factors that are associated to car thefts.

CCS CONCEPTS

• **Social and professional topics** → *Management of computing and information systems*; Systems analysis and design;

KEYWORDS

pattern extraction, car theft, Google Maps

ACM Reference Format:
Karen Oróstica and Barbara Poblete. 2019. Mining the Relationship Between Car Theft and Places of Social Interest in Santiago Chile. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.3316464

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3316464

1 INTRODUCTION

Prior work has established that crimes do not occur randomly in time and space, and that there are specific patterns for each crime class [5]. Through various analysis strategies, such as the mapping of crime hotspots, researchers have been able to identify regions with levels of differential thefts, showing that there are underlying social and spatial characteristics associated with different types of crimes. In particular, some types of places appear to be more attractive than others for certain types of crimes, for example those with a people flow and a large populations. In the case of car thefts, key factors are the number of motivated offenders, of suitable targets, and the lack of security measures, are important variables that could explain this type of crime [8]. However, car theft crimes probably depend on other of factors, such as the place where the crime takes place, the type of car that was stolen and the methods used by criminals [7]. In addition, other factors to be considered are the city or region where the crime takes place, because all cities have different social environments, population, crime statistics, and shopping mall distributions. Shopping malls, in particular, have the characteristic that they have been specifically designed to attract and seduce consumers [6]. In this study, we investigate car thefts in Chile, specifically in its most populated administrative division, the Metropolitan Region of Santiago. Santiago is the economic and social center of the country, being the most populated region, with approximately 7.1 million inhabitants [3]. In recent years, economic growth has been reported in the country, which has been reflected in the increase in the number of motor vehicles that move around the region, which reached 5.19 million in 2017 [2]. In that same year, 12,015 cases of motor vehicle thefts were reported, being one of the regions with the highest number of car thefts in Chile [1]. These statistics are alarming and show that stolen vehicle rates are a problem for Santiago. In particular, there are very few studies that attempt to analyze the problem in the region at a more quantitative level. In this work, we propose to study spatial patterns of car theft crimes in Santiago. Specifically, we focus on analyzing the influence of *places of social interest* (PSI), or of social attraction (e.g., restaurants, schools, and bars) on the concentration of *car theft points* (CTP). We study how places with high car theft frequency relate to nearby social interest points. To achieve this, we study the association between CTP and PSI through statistical and frequency analysis using a proximity-based approach. For this, we model each CTP as a vector in function of their distance to places of social interest, where each dimension represents the frequency of PSIs, for each social category, within a 200 m radius. In the same way, we modelled the PSI as a vector in function of their distance to car theft points, where the first dimension corresponds to the

frequency of CTP, and the second is the social category evaluated. These vectors will be used to compare the distribution of CTP and PSI and evaluate the location the points with high frequencies of CTP and PSIs for each social category.

In particular, our findings show that the CTP and PSI are highly associated, and also we found that there is a co-location between the most frequent points of theft and PSI in Santiago.

Our work focuses on cars theft and its relationship to places that produce social interest, with the following contributions:

(1) We introduce a methodology for identifying important points of car thefts based on the analysis of frequencies and their proximity to points of social interest.
(2) We show that there is a high correlation between CTP and PSI for all social categories in Santiago.

2 DATASET DESCRIPTION

In this section, we describe the datasets used in our study, which corresponds to car theft reports submitted to car insurance companies and to places that produce social attraction in the city according to Google Map[1].

2.1 Car theft points dataset

This dataset consists of all car thefts reports submitted by users to car insurance companies in Santiago, Chile. The dataset was provided to us by the National Association of Car Insurance Companies through a collaboration agreement[2]. This dataset consists of 23, 557 car theft records of crimes that occurred between 2010 and 2017. The original dataset contained records for all of Chile in which the location of the car theft was written in terms of street names. We geocoded car theft locations using the Google Maps Platform[3]. The geocoding process converted an address (i.e., natural language) to geographic coordinates. In particular, we found a higher concentration of stolen vehicles in the regions of Antofagasta, Valparaiso, Bio Bio and Santiago, where the latter presents the largest number of stolen vehicles.

2.2 Places of Social Interest dataset

This dataset concentrates geographical locations of places in Santiago that are likely to attract people on a day-to-day basis. We refer to these locations as PSI. To obtain the list of such places in the city we used the Google Places API [4] from the Google Maps Platform. This API provides the feature to search for places using keywords or by providing a place type. Currently, the API provides a list of 90 place types[5], such as the airport, restaurant, aquarium, art_gallery, shopping mall, among others. From these categories we selected a subset, which in general produce high people flow or might influence positively or negatively the occurrence of car thefts (e.g., shopping malls and police stations, respectively). Specifically, we used 6 categories that were: restaurant, school, shopping mall, bar, bank, and police. Finally, we filtered the remaining data points to keep only those that were within a 200 meter radius of at least one CTP in the Car theft points dataset. Overall, we remained with 127, 123 data points corresponding to places of social interest for our study. Figure 1 shows the total number of places in the dataset for each category.

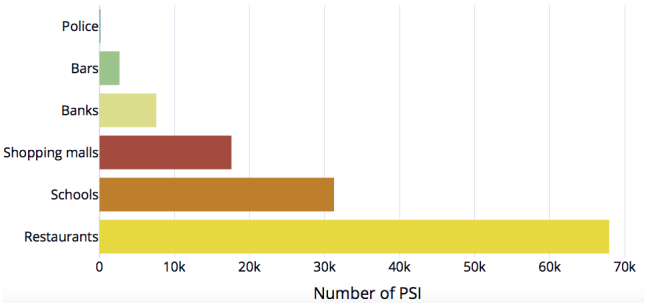


Figure 1: Total number of places of social interest in each category.

3 DATA ANALYSIS

We model each car theft point in function of their distance to places of social interest. This is, each CTP is represented as a vector in which each dimension corresponds to the frequency of social interest points in a particular category within a 200 mt radius.

$$\text{CTP}_{<lat,lon>} =< freq_{category1}, \ldots, freq_{category6} >$$

To model the places of social interest, we decided to represent each PSI as a vector, where the dimensions included to social category and the frequency of car theft points within a 200 mt radius.

$$\text{PSI}_{<lat,lon>} =< freq_{CTP}, Social_{categories} >$$

To obtain the frequencies of CTP and PSI within of specific radius we used the kd-tree algorithm implementation [4] in the R RANN package[6].

Car theft frequency analysis We want to identify geographical locations in Santiago where car thefts occurred more frequently than others places. For this, we obtained geographical locations and its number of car theft in Santiago. In Figure 2 we show geographical locations where the car thefts frequently occur in the city. We represent the points with the lowest frequency of car theft (less than 5) as light yellow dots. Geographic locations with an intermediate frequency of car theft (greater than 5 and less than 50) are displayed as yellow dots. In red, we show geographic locations with highest frequency of car theft (high than 50). we noted that geographical locations are not distributed randomly; on the contrary, we observed that there is a trend towards the northeast sector of the region, which is the wealthiest sector of the city. This result is consistent with what we expected since the data set corresponds to insured cars, which have a higher value than the average of cars. The great majority of the points have a frequency of thefts less than 5, but also there are some places with an high frequency of car thefts. We identify 4 points of car thefts with a frequency of over 50, which are located in the southeast, north, middle and East

[1] http://maps.google.com
[2] Project FONDEF project ID16I10222, CONICYT.
[3] https://developers.google.com/maps/documentation/geocoding/intro
[4] https://developers.google.com/places/web-service/intro
[5] https://developers.google.com/places/supported_types?hl=es
[6] https://cran.r-project.org/web/packages/RANN/index.html

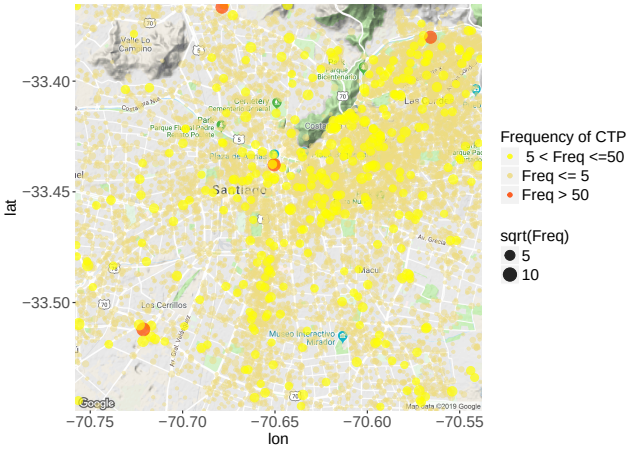


Figure 2: Frequency categories of car thefts: low (light yellow), intermediate (yellow) and high (red) in Santiago.

of the city. These geographical locations are very close to the main shopping centres in Santiago. These shopping centres correspond to *Mall Plaza Oeste*, *Mall plaza Norte*, *Plaza de Armas* and *Mall Alto las Condes*, respectively.

Identification of hotspots based in CTP and PSI vector To understand the relationship between car thefts and their proximity to places that attract large numbers of people, we characterize the car theft points (i.e, geographic locations) were near places of social interest with a high frequency of car thefts. To evaluate this idea, we created histograms for all dimensions of CTP and PSI, where we categorized each CTP and PSI into ranges according to its frequency. We selected CTP with a high frequency of PSI, for all dimension, and also, we choose PSI with a high frequency of car theft for all social categories. Then, we show the selected points on the map of Santiago. To the side of the maps, we display the corresponding histograms of CTP and PSI. The Figure 4 shows these results for police (A), restaurant (B) and school (C) categories, and the Figure 5 display the results for bank (A), bar (B) and shopping mall (C). We found that there is a high correspondence between the neighbourhoods recognized by social categories evaluated in this work and the highlighted geographical locations obtained by our method. For example, we noticed that the main shopping malls of Santiago were identified by our method as hotspots of car thefts based on the frequency of CTP and PSI. In the case of restaurant, we

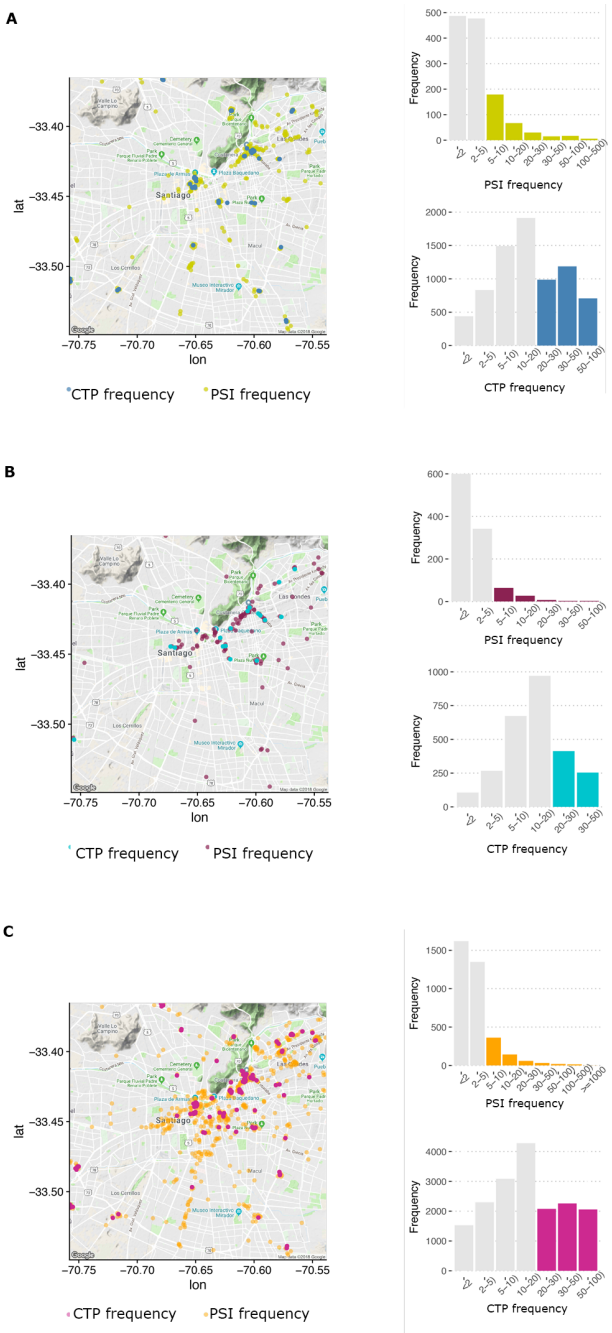


Figure 3: Geographic locations with a high number of PSI and car theft, and the distribution of these points for bank (A), bar (B) and shopping mall (C).

identify neighborhoods of recognized restaurants, such as the*Barrio Italia, Avenida Tobalaba, Barrio Brasil*, with a high frequency of CTP and PSI. The same occurs in the other social categories.

Comparison of the distribution of CTPs and PSIs We obtained the correlation between CTP and PSI agrouped by each

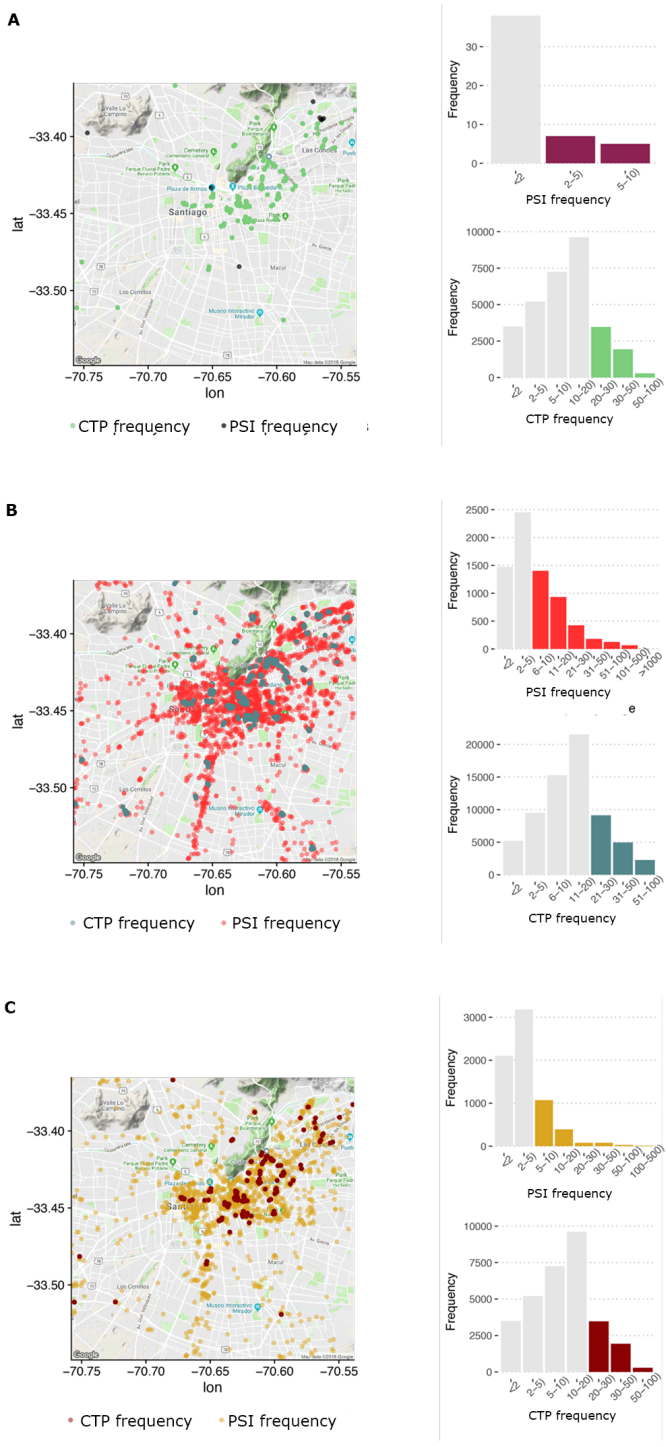


Figure 4: Geographic locations with a high number of PSI and car theft, and the distribution of these points for police (A), restaurant (B) and school (C).

districts of Santiago using Spearman method. We decided to use this administrative division because there is an important economic and social difference between them. We observed a high correlation between CTP and PSI for all social category, where Restaurant reached the highest correlation (r = 0.98) and police got the lowest correlation (r = 0.73) (See Table 1). These results support our idea of insured car thefts are associated with places attractive socially, like restaurants, schools, etc.

Table 1: Correlation between CTPs and PSI

Category of PSIs	Correlation coefficient
Restaurant	r = 0.98 (P = 2.2e-16*)
School	r = 1 (P = 2.2e-16)
Shopping mall	r = 0.97 (P = 2.2e-16*)
Bar	r = 0.97 (P = 2.2e-16*)
Bank	r = 0.96 (P = 2.2e-16*)
Police	r = 0.73 (P =1.3e-0*9)

3.1 Discussion and Conclusion

We study how the thefts of the insured car are mediated by the proximity to places of social interest. In particular, we proposed a new methodology to identify hotspot of car thefts car based on the frequency analysis using information from *car theft points* (CTP) and *Places of Social Interest* (PSI). In summary, we were to able to identify neighbourhoods or streets with a high frequency of CTP and PSI, which are neighbourhoods recognized by social activity, for example, restaurants, schools, bars, among others. In addition, it was seen that there were common hotspots between categories, for example, restaurant and shopping malls.

On the other hand, with the information provided by this new methodology, insurance companies can improve their service, it is even possible to use this information to create a system of recommendation to the user to avoid future thefts of their cars.

ACKNOWLEDGMENTS

Funding support national scholarship conicyt 21182123.

REFERENCES

[1] Carabineros de Chile. 2018. Statistical Report. http://dac.carabineros.cl/Reportes/reporte_noviembre_2018.html. [Online; accessed 22-January-2018].

[2] Instituto Nacional de Estadística de Chile. 2018. Parque de Vehículos. https://www.ine.cl/estadisticas/economicas/transporte-y-comunicaciones?categoria=Anuarios. [Online; accessed 22-January-2018].

[3] Biblioteca del Congreso Nacional. 2018. Región Metropolitana de Santiago. https://www.bcn.cl/siit/nuestropais/region13/index.htm. [Online; accessed 22-January-2018].

[4] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. 1977. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. Math. Softw.* 3, 3 (Sept. 1977), 209–226. https://doi.org/10.1145/355744.355745

[5] Jerry Ratcliffe. 2010. Crime Mapping: Spatial and Temporal Challenges. In *Handbook of Quantitative Criminology*, Alex R. Piquero and David Weisburd (Eds.). Springer New York, New York, NY, 5–24. https://doi.org/10.1007/978-0-387-77650-7_2

[6] Joel Stillerman and Rodrigo Salcedo. 2012. Transposing the Urban to the Mall: Routes, Relationships, and Resistance in Two Santiago, Chile, Shopping Centers. *Journal of Contemporary Ethnography* 41, 3 (June 2012), 309–336. https://doi.org/10.1177/0891241611434551

[7] Isaac T. Van Patten, Jennifer McKeldin-Coner, and Deana Cox. 2009. A Microspatial Analysis of Robbery: Prospective Hot Spotting in a Small City. *Crime Mapping: A Journal of Research and Practice* 1, 1 (2009), 7–32. http://ivanpatt.asp.radford.edu/Research/Microspatial%2520Analysis%2520of%2520Robbery.pdf

[8] Yuanyuan Mao, Shenzhi Dai, Jiajun Ding, Wei Zhu, Can Wang, and Xinyue Ye. 2018. Space Time Analysis of Vehicle Theft Patterns in Shanghai, China. https://www.mdpi.com/2220-9964/7/9/357/htm

NIFify: Towards Better Quality Entity Linking Datasets

Henry Rosales-Méndez
DCC, University of Chile
hrosales@dcc.uchile.cl

Aidan Hogan
IMFD; DCC, University of Chile
ahogan@dcc.uchile.cl

Barbara Poblete
IMFD; DCC, University of Chile
bpoblete@dcc.uchile.cl

ABSTRACT

The Entity Linking (EL) task identifies entity mentions in a text corpus and associates them with a corresponding unambiguous entry in a Knowledge Base. The evaluation of EL systems relies on the comparison of their results against gold standards. A common format used to represent gold standard datasets is the NLP Interchange Format (NIF), which uses RDF as a data model. However, creating gold standard datasets for EL is a time-consuming and error-prone process. In this paper we propose a tool called NIFify to help manually generate, curate, visualize and validate EL annotations; the resulting tool is useful, for example, in the creation of gold standard datasets. NIFify also serves as a benchmark tool that enables the assessment of EL results. Using the validation features of NIFify, we further explore the quality of popular EL gold standards.

CCS CONCEPTS

• **Information systems** → *Information extraction.*

KEYWORDS

Information Extraction; Entity Linking; Benchmark Dataset

ACM Reference Format:
Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2019. NIFify: Towards Better Quality Entity Linking Datasets. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316465

1 INTRODUCTION

Entity Linking (EL) involves annotating entity mentions in a text and associating them with a corresponding unambiguous identifier in a Knowledge Base (KB). EL has gained increasing attention in recent years due mainly to the availability of large KBs on the Web (e.g., Wikipedia, DBpedia, Wikidata, BabelNet) that offer unambiguous identifiers and relevant information for a wide range of entities. For instance, in the sentence **S1** *"Jackson won an award as best-selling artist of the 1980s"* an EL system targeting the DBpedia KB should identify *Jackson* as `dbr:Michael_Jackson`[1]; in this way, we know that the text speaks about a famous musician from the U.S. who is also known as the *King of Pop*. EL thus helps to build a bridge from unstructured information (text) to (semi-)structured data (KBs). Many applications then rely on EL, including semantic search, semantic annotation, text enrichment, entity summarization, relation extraction, and more besides.

Several EL systems have been proposed thus far, along with a range of gold standards for evaluation purposes (surveyed later in Table 1). However, as research on EL has continued to advance, more specialized requirements are being considered, reflecting real environments that stand to benefit from EL; such requirements include multilingualism, specific domains, noisy texts, short texts, semi-structured inputs, etc. With this diversification of requirements, traditional gold standards are not enough: novel gold standards are ideally required to reflect different contexts.

Gold standard datasets are commonly built manually by expert humans reflecting a ground truth. Early datasets were written in (varying) ad hoc formats that required special processing. Hellmann et al [6] thus proposed the NLP Interchange Format (NIF) in order to improve the interoperability of NLP tools, including EL tools. NIF is based on the RDF data model, defining a vocabulary in OWL for representing and sharing NLP-related annotations.

Despite the benefits of NIF, the creation of gold standards is still a complex, error-prone and time-consuming work; hence a number of tools have been proposed to help experts in this task. Röder el al. [17] craft three NIF datasets from texts written in English and German that were tagged manually using their own tool, but to the best of our knowledge the tool is not openly available. Looking for mistakes in datasets, Kunal et al. [9] propose guidelines to validate EL datasets, providing the EAGLET system that checks a variety of quality rules, helping experts to reduce errors; however, some important errors, such as verifying that the target of a link is not a redirect page, are not covered. On the other hand, other works have focused on standardizing the assessment process, providing benchmarking suites (e.g., GERBIL [20], Orbis [15]) that can quickly compare results for state-of-the-art EL systems against a variety of datasets. More generally, all of these NIF operations – creating, validating and performing experiments with EL datasets – have, to the best of our knowledge, been addressed as independent systems.

In this short paper, we thus describe NIFify: a tool that simultaneously supports the creation, visualization, and validation of NIF datasets, as well as the comparison of EL systems. With our tool – shown in Figure 1 – we include some functionalities not covered by previous approaches for creating, modifying and validating NIF datasets. Additionally, we allow to visualise the results of EL systems at both a sentence and document level.

2 BACKGROUND

The typical way to evaluate EL systems is through gold standard datasets, which contain text corpora and their corresponding annotations of entity mentions with respect to the identifiers of a given KB (or multiple KBs). One can then use such datasets in order to measure the quality of the output of an EL system. As more and more such datasets were proposed for EL, interoperability became

[1] Throughout, we use well-known prefixes according to http://prefix.cc

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316465

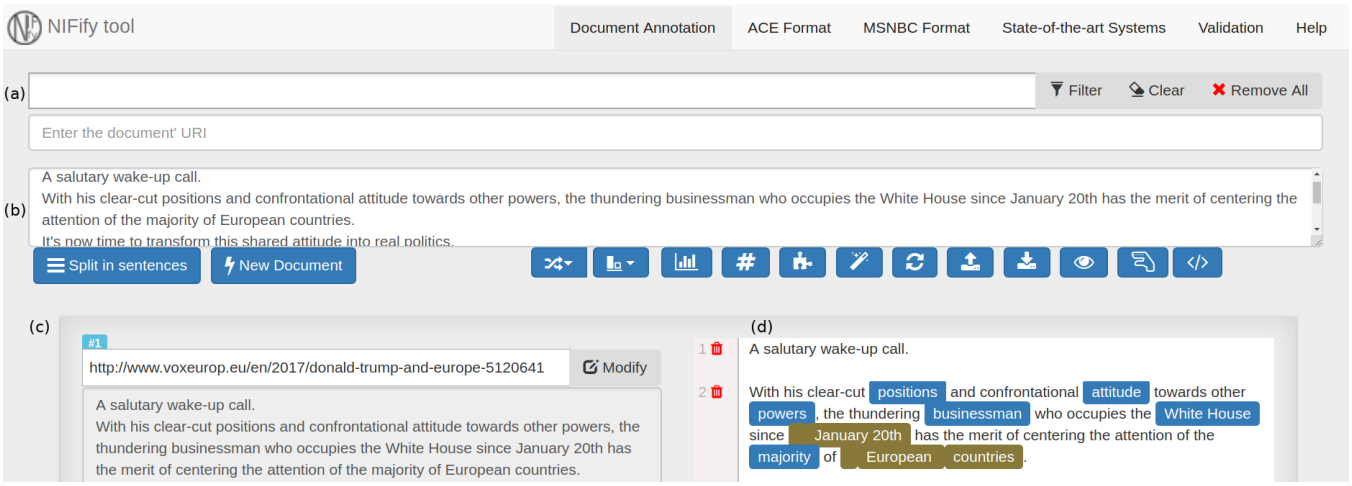


Figure 1: The main view of NIFify showing: (a) the class-reference input to filter annotations; (b) the document text input; (c) the mention identification field; and (d) the annotation visualization.

an issue: various formats were used to represent such datasets. One of the first formats proposed for EL annotation was for the MSNBC [4] dataset, which has two separate files: one a plain text file, and the other an XML file describing the annotations. This same format was followed by other authors proposing further EL datasets. e.g., ACE2004 [16], AQUAINT [16], IITB [10].

However, other EL datasets began to follow other formats. In Table 1 we list some of the most popular EL datasets in the literature along with some details of their content: whether or not they were created manually (**Mn**), whether or not the entity mentions are explicitly typed (**Typ**), and the format used. In terms of formats, many are based on XML (e.g., MSNBC [4], IITB [10], RENDEN [1], CAT [12]) or CSV (e.g., AIDA [7], SemEval [13]). However, a number also use RDF as a base data-model: Melo et al. [5] proposed Lexvo[2] as a RDF-based format and service that defines a unique URI for terms, languages, scripts, and characters from a text corpus; later, Hellmann et al. [6] the NLP Interchange Format (NIF), based on RDF, which is interoperable with a variety of NLP tools, and has been used by several recent EL datasets (e.g., N3-RSS 500 [17], Reuters 128 [17], Wes2015 [22], News-100 [17], DBpedia Abstracts [2], VoxEL [18]). Further legacy datasets were transformed to NIF, including KORE50 and DBpedia Spotlight[3].

NIF is based on RDF triples <*subject*, *predicate*, *object*> where the *subject* identifies a unit of information, such as a document, sentence, or annotation; and each *predicate*—*object* pair defines values for their properties. Figure 2 provides a brief example of a single entity annotation serialized in the Turtle syntax of RDF. The properties `nif:beginIndex` and `nif:endIndex` indicate the start and end position of the entity mention in a sentence; the targeted KB identifier is specified using the property `itsrdf:taIdentRef`; and a class can be defined with `itsrdf:taClassRef`. Other NIF properties capture metadata for other NLP tasks, such as stemming (`nif:stem`), part-of-speech tagging (`nif:oliaCategory`, `nif:lemma`), etc.

[2]http://lexvo.org/ontology; January 1st, 2019.
[3]http://apps.yovisto.com/labs/ner-benchmarks; January 1st, 2019.

Table 1: Overview of popular EL datasets; we highlight in bold those datasets that have been converted to NIF

Dataset	Mn	Typ	Format
MSNBC [4]	✗	✗	MSNBC
IITB [10]	✓	✗	IITB
AIDA/CoNLL [7]	✓	✗	AIDA
ACE2004 [16]	✗	✗	MSNBC
AQUAINT [16]	✗	✗	MSNBC
DBpedia Spotlight [11]	✓	✗	Lexvo
KORE50 [8]	✓	✗	AIDA
N3-RSS 500 [17]	✓	✗	NIF
Reuters 128 [17]	✓	✗	NIF
News-100 [17]	✓	✗	NIF
Wes2015 [22]	✓	✗	NIF
SemEval 2015 Task 13 [13]	✓	✗	SemEval
Thibaudet [1]	✗	✓	RENDEN
Bergson [1]	✗	✓	RENDEN
DBpedia Abstracts [2]	✗	✗	NIF
MEANTIME [12]	✓	✓	CAT
VoxEL [18]	✓	✗	NIF

Figure 2: NIF triples to specify the annotation of Jackson from sentence S1

```
<https://example.org/doc1#char=0,7> a nif:String,
    nif:Context, nif:Phrase, nif:RFC5147String;
    nif:anchorOf """Jackson"""^^xsd:string ;
    nif:beginIndex "0"^^xsd:nonNegativeInteger ;
    nif:endIndex "7"^^xsd:nonNegativeInteger ;
    itsrdf:taIdentRef </wiki/Michael_Jackson> .
```

3 NIF CONSTRUCTION

A number of EL datasets have either been computed from existing sources, or computed automatically. For example, DBpedia Abstracts is too large for human labeling to be feasible.[4] On the other hand, the recently proposed BENGAL tool [14] adopts a creative strategy for automatically generating gold standard datasets: rather than start with text, the authors propose to start with facts about entities from structured datasets (in RDF) and use verbalization components to convert these facts to text, recording which entities are used to generate which sentences; while this approach has the benefit of being able to generate very large and accurate gold standards, how representative the generated text is of real-world corpora depends on the quality of the verbalization component.

On the other hand, per Table 1, most datasets are constructed with manual intervention, and a number of systems have been proposed to help in this process. In previous work, we manually annotated a multilingual EL dataset called VoxEL [18], generating NIF annotations; at the start of this process, we tried to find an existing tool that would aid in the annotation process, but we found that while some systems were unavailable, others (e.g., QRTool[5]) we could not install, or did not offer features such as validation.

Addressing these limitations, we propose NIFify: an open source tool that provides end-to-end support for EL annotation, including the import of text corpora[6]; the import (including the conversion of MSNBC formats to NIF) of existing EL datasets; the addition and revision of annotations; custom tagging systems for annotations; visualizations of annotations; overlapping mentions; and finally, visualisations of the results of EL systems over the resulting dataset. The tool requires no installation and can be used either online or offline in a browser[7]. For space reasons, rather than describe all the features of NIF, we focus on two group of features of particular importance to NIFify: *validation* and *result visualization*.

4 VALIDATION

Validation is a crucial step to help human experts ensure the production of a ground truth for gold standards, and EL datasets are no exception. Legacy EL datasets have been observed to contain errors or design choices that may affect the results of evaluation [9, 19, 21]; furthermore, target KBs may evolve, rendering some links obsolete.

Erp et al. [21], analyze characteristics of seven EL datasets and find biases introduced by the decisions taken in the annotation process; they highlight the need for a more standard creation of datasets. Jha et al [9] propose a set of validation rules and propose the EAGLET system to check these rules when constructing EL datasets; however, these rules are sometimes dogmatic, considering, for example, overlapping mentions to be errors when they are considered valid by other definitions [19]; furthermore, EAGLET requires execution on a command-line to highlight errors in the visualization, rather than being supported by the interface.

NIFify allows for detecting possible errors present in terms of the mentions and the identifiers to which they are linked; specifically, the following rules are checked:

[4]Details of the annotation process are not provided, but we assume it uses links already present in the corresponding Wikipedia texts.
[5]https://github.com/dice-group/QRTool; January 1st, 2019
[6]https://users.dcc.uchile.cl/~hrosales/MSNBC_ACE2004_to_NIF.html; Jan. 1st, 2019
[7]https://github.com/henryrosalesmendez/NIFify_v2; January 1st, 2019

Table 2: Errors found in current NIF datasets; the last dataset was labeled by us

Dataset	SE	LE	FE	CE
DBpedia Spotlight	8	23	4	–
N3-RSS 500	1	34	–	–
Reuters 128	4	71	–	–
News-100	9	1515	–	–
Wes2015	–	609	–	–
VoxEL	–	8	–	–

- Spelling Error (SE): Mentions should neither start nor end in the middle of a word.
- Link Error (LE): When linking to Wikipedia or DBpedia, identifiers should be the URLs/IRIs corresponding to an unambiguous, non-redirect page on Wikipedia.
- Format Error (FE): We check the consistency of the NIF representation with two sub-rules:
  - Annotations are typically assigned a subject IRI of the form `http://example.org#char=x,y`, where x and y should correspond with the values given for `nif:beginIndex` and `nif:endIndex` respectively.
  - The substring identified by these positions should correspond with that denoted by the `nif:anchorOf` property.
- Category Error (CR): For those datasets with classes specified by the predicate `itsrdf:taClassRef`, NIFify allows the specification of custom rules in order to detect inconsistencies in the annotation classes. For example, the classes `dbo:Person` and `dbo:Event` should not be present on the same annotation as they are disjoint: an entity is typically not a person and an event at the same time.

NIFify then encodes rules to detect these errors and thus validate EL datasets. In order to test the prevalence of these errors in existing datasets, we ran NIFify's validation over EL datasets currently available in the NIF format (excluding those that we converted ourselves to NIF – MSNBC and ACE2004 – since we resolve such errors as part of the conversion). In Table 2, we show the results of this validation process, where we can observe that all datasets considered contain errors of at least one type.

In the majority of the cases, SE errors are introduced in the construction of the dataset with the addition of characters that do not belong to the mention, or on the contrary, leaving out part of a word that completes a mention; for example, in the DBpedia Spotlight dataset, the URI `wiki:Man` is associated with the three characters of the world *performance*. Other SE errors contained in the datasets involve missing spaces between words.

The most frequent type of error encountered in the NIF dataset was LE: this is mainly due to the fact that KBs are constantly evolving, which may affect link consistency. For example, in Wikipedia, pages about specific entities may become disambiguation pages, or redirects to other pages. Such changes explain why our own dataset (VoxEL, created using NIFify) contains such errors: the external KB has evolved since its creation. The News-100 and Wes2015 contain a large number of LE errors beyond what can be explained by the KB changing: for example, in the Wes2015 dataset, 520 of its LE

errors correspond to redirect pages, 48 to disambiguation pages, while the rest do not point to valid pages.

Finally, the only dataset we found with FE-type errors was DBpedia Spotlight, which had problems with its NIF representation. On the other hand, we did not find any errors of type CE.

We have published all errors found online for reference.[8] We conclude that most of the validation features of NIFify can help to improve the quality of EL datasets, including to find problems caused by the evolution of a KB over time.

5 RESULT VISUALIZATION

Once an EL dataset has been generated, the next step is to evaluate and compare EL systems using the dataset. A number of systems have been proposed to help evaluate and compare EL systems. Cornolti et al. [3] proposed the BAT framework, which they used to compare five EL systems over five datasets. Along similar lines, Usbeck et al. proposed GERBIL [20], which extends the systems and (NIF) datasets supported. However, both frameworks produce comparative metrics, rather than visualizing the actual output of the EL tool(s). Another EL benchmark framework called Orbis [15] was recently proposed that includes visualization of systems' responses; however, Orbis is not available in the provided URL.[9].

Given that there is no clear definition on what EL systems should link [19], we argue that metrics like precision and recall may not tell the full story, and that results may be due not only to the quality of the output produced by an EL system, but also whether or not it targets the same types of entities as labeled in the dataset. Comparing EL results with the ground truth labeled in a dataset under construction/revision may even lead to changes in the dataset.[10] Hence with NIFify we propose a benchmark framework to visualize the results of EL systems over the NIF dataset, highlighting both *true positives* or *false positives*, which allows a more qualitative assessment of both a given EL tool and an EL dataset, possibly in the context of a given application. Additionally, NIFify can be used to demo EL systems, offering a visual, friendly user interface.

6 CONCLUSION

In this short paper, we describe the NIFify system, which aims to address a number of shortcomings of existing tools for generating EL datasets and evaluating EL tools: in particular, NIFify simultaneously supports the creation, visualization, and validation of NIF datasets, as well as the comparison of EL systems. We first discussed some extensions to the NIF format to support mentions having multiple possible identifiers annotated with different types. We then provided a summary of the main features of NIFify for generating EL gold standard datasets, before focusing on features relating to validation, showing that existing EL datasets exhibit errors detectable by the tool, detecting a total of 2,321 errors across six datasets; we publish these errors online for reference: https://users.dcc.uchile.cl/~hrosales/dataset_errors.html. Finally, we discuss the importance of features for visualizing the results produced by an EL system, which are further implemented in the NIFify tool. A demo of the tool is available at https://users.dcc.uchile.cl/~hrosales/NIFify_v2.html

[8] https://users.dcc.uchile.cl/~hrosales/dataset_errors.html; January 1st, 2019.
[9] https://github.com/htwchur; January 1st, 2019.
[10] Of course, we urge caution to ensure that bias is not introduced by adapting a dataset to suit a subset of tools evaluated.

7 ACKNOWLEDGMENTS

The work of Henry Rosales-Méndez was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017. The work was also supported by the Millennium Institute for Foundational Research on Data (IMFD) and by Fondecyt Grant No. 1181896.

REFERENCES

[1] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: named entity linking in digital literary editions using linked data sets. *CSIMQ* 7 (2016), 60–80.
[2] Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus.. In *LREC*.
[3] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *World Wide Web Conference*.
[4] Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *EMNLP-CoNLL* (2007), 708.
[5] G. de Melo and G. Weikum. 2008. Language as a foundation of the Semantic Web. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference*.
[6] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP Using Linked Data. In *ISWC*. 98–113.
[7] Johannes Hoffart and et al. 2011. Robust disambiguation of named entities in text. In *EMNLP*. ACL, 782–792.
[8] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *CIKM*. 545–554.
[9] Kunal Jha, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. All that Glitters Is Not Gold - Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking. In *ESWC*. 305–320.
[10] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *SIGKDD*. 457–466.
[11] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*. ACM, 1–8.
[12] A.L. Minard and et al. 2016. MEANTIME, the NewsReader multilingual event and time corpus. (2016).
[13] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking.. In *SemEval@ NAACL-HLT*. 288–297.
[14] Axel-Cyrille Ngonga Ngomo, Michael Röder, Diego Moussallem, Ricardo Usbeck, and René Speck. 2018. BENGAL: An Automatic Benchmark Generator for Entity Recognition and Linking. In *Proceedings of the 11th International Conference on Natural Language Generation*. 339–349.
[15] Fabian Odoni, Philipp Kuntschik, Adrian M. P. Brasoveanu, and Albert Weichselbraun. 2018. On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance. In *SEMANTICS*. 33–42.
[16] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *NAACL-HLT*. 1375–1384.
[17] Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In *LREC*. 3529–3533.
[18] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In *ISWC*. 170–186.
[19] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What Should Entity Linking link?. In *AMW*.
[20] Ricardo Usbeck and et al. 2015. GERBIL: General Entity Annotator Benchmarking Framework. In *WWW*. 1133–1143.
[21] Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC*.
[22] Jörg Waitelonis, Claudia Exeler, and Harald Sack. 2015. Linked data enabled generalized vector space model to improve document retrieval. In *NLP & DBpedia @ ISWC*.

After Brazil's General Data Protection Law: Authorization in Decentralized Web Applications

Jefferson O. Silva
silvajo@pucsp.br
Pontifical Catholic University of São Paulo
São Paulo, Brazil

Newton Calegari
newton@nic.br
Brazilian Network Information Center - NIC.br
São Paulo, Brazil

Eduardo S. Gomes
egomes@pucsp.br
Pontifical Catholic University of São Paulo
São Paulo, Brazil

ABSTRACT

Decentralized web applications do not offer fine-grained access controls to users' data, which potentially creates openings for data breaches. For software companies that need to comply with Brazil's General Data Protection Law (LGPD), data breaches not only might harm application users but also could expose the companies to hefty fines. In this context, engineering fine-grained authorization controls (that comply with the LGPD) to decentralized web application requires creating audit trails, possibly in the source code. Although the literature offers some solutions, they are scattered. We present Esfinge Guardian, an authorization framework that completely separates authorization from other concerns, which increases compliance with the LGPD. We conclude the work with a brief discussion.

CCS CONCEPTS

• **Software and its engineering → Software design engineering**.

KEYWORDS

Access control; Decentralized Web Applications; Frameworks; Guardian; Solid

ACM Reference Format:
Jefferson O. Silva, Newton Calegari, and Eduardo S. Gomes. 2019. After Brazil's General Data Protection Law: Authorization in Decentralized Web Applications. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316461

1 INTRODUCTION

With the approval of the Brazilian General Data Protection Law (LGPD, Portuguese acronym) [7], several software companies may need to redesign the applications that handle the personal data of Brazilian citizens. The LGPD considers personal any data that directly or indirectly lead to the identification of a user [7]. Neglecting the LGPD requirements could mean incurring in fines up to 2% of companies' global revenue [7].

The LGPD sets compliance requirements on the companies in charge of making decisions about the data processing (i.e., data controllers) and the companies that process personal data in the name of data controllers (i.e., data processors) [7]. Besides, the LGPD states that, in some cases, data controllers and data processors may be held liable, especially in cases where data breaches are harmful to users [7].

To avoid being classified as either data processors or data controllers (to avoid sanctions), some companies may redesign applications as decentralized web applications. In the context of this research, an application is considered decentralized when it does not hold users' data. Berners-Lee and colleagues [12] proposed a platform called Solid (derived from "**So**cial **li**nked **d**ata"), which can be described as a set of principles, conventions, and tools for building decentralized web applications. Solid is based on the principle that users should have full ownership of their data, which are stored in Web-accessible personal online datastores (pods) [12]. Pods are independent of web applications. For obtaining services, users need to authorize web applications to access their pods explicitly, by classifying web applications as trusted.

Using Solid alone leaves users solely responsible for controlling access to protected resources, which may not be enough to comply with the LGPD. The LGPD requirement of data governance (see Art. 50, Par. 2 in [7]) states that, among other things, companies should establish adequate policies to protect users' data. Nevertheless, in Solid web applications, a user would not have the means to prevent unauthorized access to their data, after classifying a web application as trusted. For example, a hospital web application may have a sensitive operation that reads personal data from patient's pods that should be accessible only by designated doctors. A violation of this access control policy would configure a data breach, in which case the hospital might still be held liable. Also, the liability risk might create the need for audits, in which case it would be necessary that the hospital demonstrated that it possesses appropriate controls, possibly directly in source code.

This context indicates that it is necessary to engineer fine-grained authorization controls, without loosing the simplicity required in auditing the source code. Thus, we establish the following research question (RQ).

RQ: How to design fine-grained authorization controls to decentralized web applications that comply with the LGPD requirement of data governance?

The answer to our RQ may help companies to increase compliance with the LGPD by employing several software engineering techniques, which are implemented in Esfinge Guardian. With this paper, we contribute to the literature in at least two ways. First, by showing that the decentralization of a web application may not be enough for companies to avoid liability issues. Second, by pointing

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316461

the need for more research on how the use of Esfinge Guardian (and others) could increase compliance with the LGPD (and other regulations).

This work is organized as follows. In Chapter 2, we offer some background. In Chapter 3, we present Esfinge Guardian. In Chapter 4, we offer a case example. In Chapter 5, we present some related works. We conclude with a brief discussion in Chapter 6.

2 BACKGROUND

In this section, we offer some background for the understanding of the research problem domain.

2.1 Brazil's General Data Protection Law

The Brazil's General Data Protection Law (LGPD) is based on the General Data Protection Regulation (GDPR),[1] which aims at protecting the personal data of EU individuals. In total, around 120 countries adopt comprehensive privacy laws and regulations to protect personal data held by private and public bodies [2]. The LGPD applies to any individual or legal entity (public or private) with personal data processing activities that: are carried out in Brazil; offer or supply goods or services in Brazil or relate to individuals located in Brazil, and; involve personal data collected in Brazil.

2.2 Decentralized Solid Web Applications

Traditional web applications (e.g., Facebook, CRMs, and hospital applications) rely on private APIs, exclusive access control mechanisms, and dedicated data storage sytems. Because users cannot move personal data to other platforms, these web applications become "data silos." We refer to these web applications as centralized [16].

Solid is a platform that supports decentralized web applications, by relying on open standards and semantic web technologies [3]. In the Solid platform, applications run in a browser or as mobile applications, while users data are stored in pods [4]. Although pods can be stored locally, they typically are stored in dedicated servers, which manage data according to the Linked Data Platform recommendation, enabling it to manipulate data items through HTTP requests [15]. Solid servers are application-agnostic and can deal with both structured and unstructured data. Structured data is represented using RDF, a Semantic Web standard [4, 10]. Application development based on Solid platform supports portability and interoperability, so applications can be seen as an interface that works with distributed data in multiple pod server implementations.

Identity in the Solid context is based on WebID, which allows agents (e.g., a person, an organization) to create their identities using global unique identifiers - HTTP URIs [12]. A WebID is an open and decentralized identification mechanism being developed by a W3C community group.[2]

2.3 Authorization in Solid

Access control is typically split into two distinct procedures: authentication, and authorization. While authentication is concerned with determining whether an agent (e.g., user, group) is whom it claims to be, authorization is responsible for verifying if the agent is allowed to access a protected resource (e.g., document) or operation (e.g., read, write, append). In this research, we focus on authorization.

Solid uses the Web Access Control (WAC) specification for authorizing the access to protected resources. According to the specification, WAC has the following key features:

(1) The resources are identified by URLs and can refer to any web documents or resources;
(2) It is declarative – access control policies are written in regular web documents;
(3) Users and groups are also identified by URLs (WebIDs);
(4) It is cross-domain – all of its components, such as resources, agent WebIDs, and even the documents containing the access control policies, can potentially reside on separate domains;

```
# Contents of https://alice.databox.me/docs/file1.acl
@prefix  acl:  <http://www.w3.org/ns/auth/acl#>  .

<#authorization1>
    a             acl:Authorization;
    acl:agent     <https://alice.databox.me/profile/card#me>;  # Alice's WebID
    acl:accessTo  <https://alice.databox.me/docs/file1>;
    acl:mode      acl:Read,
                  acl:Write,
                  acl:Control.
```

Listing 1: Example WAC Document

Listing 1 shows an example of a WAC document that specifies that Alice (as identified by her WebID `https://alice.databox.me/profile/card#me`) has full access (read, write, and control) to one of her web resources, located at `https://alice.databox.me/docs/file1`.

Similarly, it is possible to give access to a group of agents using the `acl:agentGroup` predicate [1]. A group is a collection of members (or WebIDs) that needs to be specified in a different file. Moreover, it is possible to give access to all agents (public access) or yet to all authenticated agents. Besides, it is also possible to classify web applications as trusted. Furthermore, not every document needs its own individual access control list file. Rather, it is possible to authorize a container, which is a web location that contain multiple resources. As mentioned, for controlling the access to the data in their pods, users need to specify WAC documents. The set of authorization modes that WAC access control systems offer does not allow the specification of fine-grained authorizations, required for decentralized web applications that needed to comply with the LGPD.

3 ESFINGE GUARDIAN

In this section, we present the Esfinge Guardian[3] framework. Essentially, the Esfinge Guardian's role is to intercept calls to protected operations. Figure 1 depicts a hypothetical interception. As an example, consider a protected operation `debit()`, which should only be executed by the account owner. Esfinge Guardian would intercept the call to `debit()`, and decide if the caller is authorized to perform the operation.

[1] http://data.consilium.europa.eu/doc/document/ST-9565-2015-INIT/en/pdf
[2] https://www.w3.org/community/webid/
[3] https://github.com/EsfingeFramework/guardian

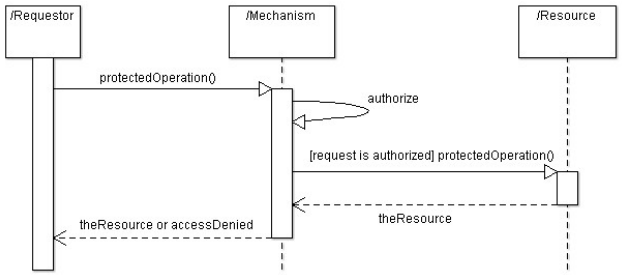


Figure 1: A conceptualization of the interception mechanism (Silva et al. [13])

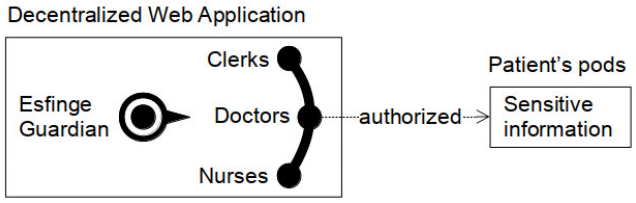


Figure 2: Esfinge Guardian representation in a decentralized web application

Additionally, Figure 2 depicts Esfinge Guardian in a hypothetical web decentralized hospital application. As depicted, Esfinge Guardian contains the authorization logic that will only allow authorized HTTP requests to a patient's pods. Besides, while Esfinge Guardian is independent of decentralized web applications (and consequentially of Solid), employing the framework may help these applications increase compliance with LGPD. We stress that although Figure 2 depicts Esfinge Guardian authorizing access to pods based on roles, it can authorize in finer-grained levels.

Esfinge Guardian is composed of eight elements (see Guerra et al. [9] and Silva et al.[13] for in-depth explanations). Figure 3 depicts the relationship between the elements in UML.

`AuthorizationContext`. It is the central entity that holds all the information required for an authorization, which includes the data for the subject, resource, and environment. That means all other entities should provide `AuthorizationContext` with enough information for authorization to occur.

`GuardianInterceptor`. Ideally, the user must be able to indicate what operations should be protected and be oblivious of all other things. A request to a protected operation must be intercepted transparently, not directly called. `GuardianInterceptor` is the entity responsible for abstracting the different existing interception technologies such as aspect-orientation, CGLib, and dynamic proxy.

`Invoker`. The original request to the protected operation on the resource is intercepted; therefore the mechanism must be in charge of replicating the request to the resource if the access is granted. `Invoker` is an entity with the ability to mimic the operation performed by the subject on a protected resource. In the Esfinge Guardian framework, this element can execute methods; however, it is important to note that it is just one of the possibilities since the architectural model is general.

One additional feature is that `Invoker` is responsible for determining when the authorization logic is performed. In many cases, enforcing the authorization logic only makes sense after the protected operation is performed. For instance, consider the case when the operation retrieves a collection, and the authorization rule requires iterating it in order to verify if the subject can indeed access all of its items. In this manner, there should be a way for configuring the precise moment that the authorization should take place.

`Populator`. It is the entity that contains the data extraction logic for authorization. Information for authorization can be anywhere such as databases, files, shared variables, user session, arguments, and the Internet. For this reason, `Populator` is an entity that knows how to obtain information from all these places. There can be zero or more `Populators` in the application; each one specialized in obtaining a different type of information from a different place.

`PopulatorProcessor`. The entity that gathers and executes all defined `Populators` in the application.

`Authorizer`. This is entity that implements the logic of the access control policy and may use information stored in `AuthorizationContext` if necessary. There must be at least one `Authorizer`. Every `Authorizer` respond in the form of a "yes" or "no;" however, it must be possible to include other response types such as "Indeterminate."

`AuthorizerProcessor`. It is the entity that contains the combining algorithm for all the `Authorizers` defined in the application.

`AuthorizationMetada`. This is an entity that indicates which resources—or their operations—must be intercepted by the authorization mechanism. A requirement is that this element must be of metadata type so that it can be used declaratively. Esfinge Guardian uses Java annotations as the implementation of this element; however, it can be considered a general marking element that is independent of a specific technology.

4 RELATED WORKS

This research combines topics usually addressed separately. We are not aware of any research that addresses the LGPD from a software engineering perspective. Alternatively, some researchers studied approaches that map GDPR principles to software design. For example, Danezis et al. [5] provided an inventory of privacy design strategies and technical building blocks of various degrees of maturity from research and development. Koops and Leenes [11] discussed whether the GDPR Privacy-by-Design principle entails hard-coding privacy requirements into applications. On decentralized web applications, Berners-Lee [3] introduced the concept of decentralized application architectures. Smith et al. [14] introduced a functional decentralized application called WebBox, while Dodson et al. [6] presented Musubi, a disintermediated interactive social feeds application for mobile devices. Concerning metadata-based frameworks such as Esfinge Guardian, Guerra et al. [8] investigated metadata usage in existing frameworks and documented recurrent solutions as architectural patterns.

5 DISCUSSION AND CONCLUSION

The LGPD requires companies to adopt a comprehensive data governance approach, including data profiling, data lineage, data masking, test-data management, and data archives. Also, specialized professionals are required to design and handle personal data. In

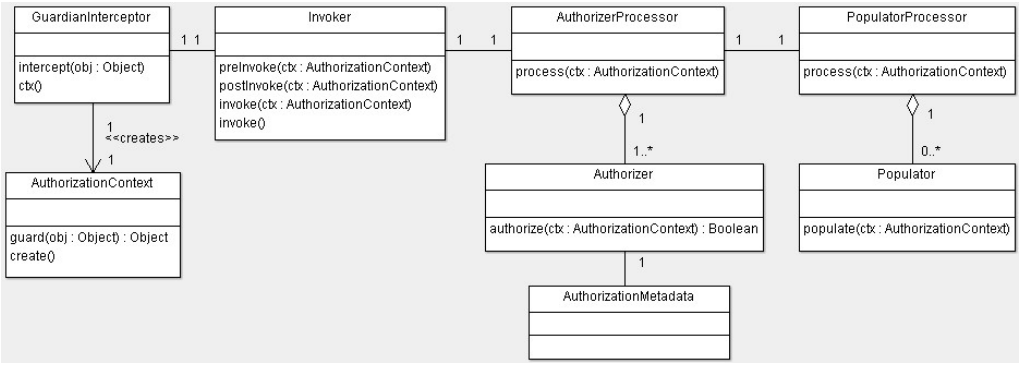


Figure 3: The Esfinge Guardian Class Diagram (Silva et al.[13])

this work, we show how Esfinge Guardian can be used to manage authorizations in decentralized web applications to increase compliance with the LGPD's data governance requirements. Besides the examples we offered, Esfinge Guardian could be used to anonymize personal data, filtering information that could lead to users identification.

ACKNOWLEDGMENTS

To the Brazilian Network Information Center (NIC.br) and the Web Technologies Study Center (Ceweb.br) for supporting this research.

REFERENCES

[1] [n. d.]. Basic Access Control ontology. https://www.w3.org/ns/auth/acl{#}
[2] David Banisar. 2011. Data Protection Laws Around the World Map. *SSRN Electronic Journal* (2011).
[3] Tim Berners-Lee. 2009. Socially Aware Cloud Storage. https://www.w3.org/DesignIssues/CloudStorage.html
[4] David Berners-Lee, Tim and Chen, Yuhsin and Chilton, Lydia and Connolly, Daniel and Dhanaraj, Ruth and Hollenbach, James and Lerer, Adam and Sheets. 2006. Tabulator: Exploring and Analyzing linked data on the Semantic Web. In *Proceedings of the 3textsuperscriptrd International Semantic Web User Interaction Workshop.*
[5] George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Metayer, Rodica Tirtea, and Stefan Schiffner. 2015. Privacy and Data Protection by Design - from policy to engineering. December (jan 2015).
[6] Ben Dodson, Ian Vo, T.J. Purtell, Aemon Cannon, and Monica Lam. 2012. Musubi: Disintermediated Interactive Social Feeds for Mobile Devices. *Proceedings of the 21st international conference on World Wide Web - WWW '12* (2012), 211.
[7] Brazilian Government. 2018. General Data Protection Law. http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm
[8] Eduardo Guerra, Clovis Fernandes, and Fábio Fagundes Silveira. 2010. Architectural patterns for metadata-based frameworks usage. *Proceedings of the 17th Conference on Pattern Languages of Programs - PLOP '10* October 2010 (2010), 1–25.
[9] E M Guerra, J O Silva, and C T Fernandes. 2015. A Modularity and Extensibility Analysis on Authorization Frameworks. 2, 1 (2015).
[10] James Hollenbach, Joe Presbrey, and Tim Berners-Lee. 2009. Using RDF metadata to enable access control on the social semantic web. *CEUR Workshop Proceedings* 514 (2009).
[11] Bert-Jaap Koops and Ronald Leenes. 2014. Privacy regulation cannot be hardcoded. A critical comment on the âĂŸprivacy by design' provision in data-protection law. *International Review of Law, Computers & Technology* 28, 2 (may 2014), 159–171.
[12] Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. [n. d.]. Solid: A Platform for Decentralized Social Applications Based on Linked Data. ([n. d.]).
[13] J.O. Silva, E.M. Guerra, and C.T. Fernandes. 2013. An extensible and decoupled architectural model for authorization frameworks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7974 LNCS.
[14] Daniel Alexander Smith, Max Van Kleek, Oshani Seneviratne, Monica schraefel, Alexandre Bertails, Tim Berners-Lee, Wendy Hall, and Nigel Shadbolt. 2012. WebBox: Supporting Decentralised and Privacy-respecting Micro-sharing with Existing Web Standards. In *21st International World Wide Web Conference.*
[15] Ashok Malhotra Steve Speicher, John Arwe. 2015. Linked Data Platform 1.0. W3C Recommendation. https://w3.org/TR/ldp/
[16] Max Van Kleek, Daniel Smith, Nigel Shadbolt, and M.c. Schraefel. 2012. A decentralized architecture for consolidating personal information ecosystems: The WebBox. *Pim 2012* (2012).

© Companion Proceedings of the 2019 World Wide Web Conference, published under Creative Commons CC-BY 4.0 License. DOI - https://doi.org/10.1145/3308560.3317076

The FAIR TRADE Framework for Assessing Decentralised Data Solutions

John Domingue

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, john.domingue@open.-ac.uk

Allan Third

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, allan.third@open.ac.uk

Manoharan Ramachandran

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, manoharan.ramachand-ran@open.ac.uk

ABSTRACT

Decentralised data solutions bring their own sets of capabilities, requirements and issues not necessarily present in centralised solutions. In order to compare the properties of different approaches or tools for management of decentralised data, it is important to have a common evaluation framework. We present a set of dimensions relevant to data management in decentralised contexts and use them to define principles extending the FAIR framework, initially developed for open research data. By characterising a range of different data solutions or approaches by how TRusted, Autonomous, Distributed and dEcentralised, in addition to how Findable, Accessible, Interoperable and Reusable, they are, we show that our FAIR TRADE framework is useful for describing and evaluating the management of decentralised data solutions, and aim to contribute to the development of best practice in a developing field.

CCS CONCEPTS

• Information systems~Distributed storage • Information systems~Resource Description Framework (RDF) • Security and privacy~Trust frameworks • Computing methodologies~Model verification and validation

KEYWORDS

Distributed Data, Evaluation, Open Data, Semantic Blockchain

1 Introduction

We now live in a world where a small number of large technology companies hold a significant amount of personal data. For example, the Guardian journalist Dylan Curran (Curran 2018) found that Google had over 5.5GB of personal data on him including everywhere he had been, and all of his Google and YouTube searches (including deleted searches). His analogous figure for Facebook was over 600MB of data. As highlighted in a number of recent scandals in the press, such as those associated with the misuse of data by Facebook and the now defunct company Cambridge Analytica (Guardian 2015), over-centralisation can cause problems for individuals as well as negatively affect democracy and national culture. The current situation has resulted in, according to the Web's inventor, "producing—with no deliberate action of the people who designed the platform—a large-scale emergent phenomenon which is antihuman." (Lee 2018a). There is also now a growing recognition of the social problems caused when centralised data systems are impoverished with respect to disadvantaged citizens, who are in effect 'data poor'. A report by the US Government's Consumer Financial Protection Bureau (Bureau 2015) found that 26 million US consumers (11% of the adult population) are 'credit invisible' and a further 19 million (8.3% of the adult population)

had credit data that was unscorable. In the UK, a lack of citizenship data led to 63 British citizens being deported (Grierson 2018) and up to 57,000 UK citizens could lose rights to homes, jobs, social benefits, UK National Health Service treatment, or be threatened with deportation in the near future (Fact 2018) – a situation which has been blamed on the combination of poorly-managed centralized data and a lack of official recognition of citizen-held data.

One type of approach to resolving the above problem has been through decentralising the data. In particular, replacing central data controllers with either the user herself, or with peer communities. Giving users control over their data and how it is processed is seen by a growing number of researchers and developers as the best mechanism to empower users and give us the 'internet we want'. The best known of these is the distributed ledger (Walport 2015) an architecture for maintaining a peer network of verified data which underpins the Bitcoin cryptocurrency. Within the Linked Data (LD) arena, Solid (Lee 2018b) provides users with their own data pods, and Linked Data Fragments (LDF) (Verborgh et al. 2016) enables data processing to be split between servers and clients. We have begun exploring the use of combined blockchain/Linked Data architectures within our own LinkChains work (Third and Domingue 2017).

Decentralisation has the potential to address problems such as those described above, but only when implemented in accordance with sound principles of decentralised data management. The problem we see for the research community is that there is currently no way of comparing decentralised data solutions as they vary greatly and, as far as we know, no comparison framework or standard exists. Our aim in this paper is to address this issue by presenting our FAIR TRADE framework for assessing decentralised data solutions, based on relevant dimensions for decentralization, and clear principles for best practice. Our framework builds upon the FAIR data principles (Wilkinson et al. 2016) which outline how scientific data should be managed, and adds dimensions related to decentralised data. Note that, just as the FAIR principles are focused on the *management* of data, so too are the TRADE principles we outline. More detailed dimensions of evaluation relating to *use* of data systems, such as performance, are not considered here.

The rest of this paper is organised as follows. In section 2 Background and Related Work, we first discuss the meanings of key terms and describe the FAIR data principles, and follow by describing a number of decentralised and distributed approaches to data management including blockchains, Solid and Linked Data Fragments, and our own decentralised approach to data management, LinkChains. Section 3 then describes our FAIR TRADE framework for assessing decentralised data approaches, before in section 5 testing our framework through the evaluation of six different data platforms. In the final section, we summarise and conclude our contributions.

2 Background and Related Work

2.1 Decentralisation and Distribution

There are a number of different ways in which data can be "spread out" across multiple locations. Each location could contain full copies of the same dataset, or different locations may contain different data. Independently, data at each location may be controlled or coordinated by a single central source, or different locations may be independent. In this paper, we refer to the former as the *distribution* of data, and to the latter as *decentralization* of data. The general theme in the literature appears to be loosely that distribution refers to location and decentralization to independence, but there is no clear consensus on definitions – e.g., Asano et al. (Asano et al. 2018) define decentralization to be when "data are maintained in different sites with autonomous storage and computation capabilities", which would make decentralised data systems a subset of distributed ones as defined by, e.g., Öszu & Valduriez (Özsu and Valduriez 2011) ("a number of autonomous processing elements […] interconnected by a computer network and that cooperate in performing their assigned tasks"). Interpreting these terms independently in the way we do permits greater expressive power, giving the option to describe all four combinations of decentralised or not vs. distributed or not.

Motivations for data distribution include redundancy, to prolong the lifespan of data and preserve it from accidental loss or malevolent attacks, as well as potentially spreading the load of data access beyond a single point of failure. Distribution of data need not be limited to simply storage and copying, but could, for example, also include querying: queries can be computationally expensive, and a solution including distributed query evaluation has the potential for better performance or sustainability than with non-distributed querying. Motivations for decentralised data systems are varied, whether a political wish to increase user control, or a practical need to use independent datasets spread across a network (e.g., the federated Linked Data model).

2.2 Autonomy and Identity

When the control of data is not restricted to those who have control over those locations in which it is stored (and, indeed, restricted *from* them), and is instead kept in the hands of particular individuals, e.g., the data owner, a data solution can be said to support *autonomy* of data. This is independent of decentralization and distribution; a fully centralized and non-distributed data system may nonetheless support autonomy if, for example, all data and associated metadata were to be encrypted under the data owners' control. There are a number of aspects of control which could be considered under data autonomy. Control over read/append/modify actions means that an individual can decide how much, if any, of their data is exposed, shared, added to, or changed, for how long and for what purpose. Particularly with personal data, control over usage and analysis may limit some of the recent abuses of data which we highlighted above. The format, metadata, and distribution of a dataset can also be controlled. Autonomy in this sense is an individual-focused concept, relative to particular users. As such, *identity* of users is important to analyzing it.

2.3 Trust

Trust in general, of course, is a broad term and includes aspects which are very difficult or impossible to address in technological systems. For example, a system cannot guarantee that data (e.g., about educational qualifications) was not fabricated prior to publication. Other aspects of trust are also covered by other principles or concepts. For example, trust that data is solely under the appropriate individual or institutional control falls under autonomy; integrity in the sense of conformance to a given schema falls under the Reusable principle of the FAIR standard, and so on. But there are some specific aspects of trust independent of those principles, and which specifically relate to decentralised data – in particular, properties relating to provenance metadata and integrity in the sense of content persistence over time. Specifically, what degree of assurance can a system provide that the publisher of a piece of data and the context of publication (e.g., date and time) can be correctly identified, and that claims based on the data are accurately based in the data originally published? In the centralised case, these are often guaranteed by the central data store, and trust in these aspects reduces to trust in the central store or its owners. Without a regulatable central authority to appeal to, these are important aspects to consider independently. (Schneider and Trustworthiness 1999) provides a high-level overview of trust in networked environments.

2.4 FAIR Data Principles

The FAIR principles emerged from a workshop held in Leiden in 2014 in recognition of the fact that the scientific community had not paid enough attention to the way digital objects including data are managed (Wilkinson et al. 2016). In particular, a need was recognised to improve the infrastructure associated with the use and reuse of scholarly data. A community of researchers, librarians, publishers and funders setup a group FORCE11 (Future of Research Communications and e-Scholarship) which agreed a set of minimal principles for the stewardship of digital scholarly artifacts. The resulting four FAIR principles are:

- ***Findable*** - data given globally unique persistent identifiers and are described via rich metadata. Metadata also has unique persistent identifiers.
- ***Accessible*** - identifiers can be used to retrieve data and metadata via a standard communications protocol which includes authentication and authorization when necessary.

- ***Interoperable*** **-** data and metadata are represented in a formal, accessible and applicable knowledge representation language. Vocabularies are also used to enable interoperability for both data and metadata. Qualified references are contained in (meta)data to point to other (meta)data.
- ***Reusable*** - data and metadata are described with relevant attributes, clear licenses and detailed provenance information.

The above principles have proved popular with the scientific community and have spawned a number of initiatives such as GO FAIR[1] which links the principles to the European Open Science Cloud. The FORCE11 community now has over 2,600 active members.[2] The FAIR principles are not inherently specific to scientific data management, and, as Mons, et al., (Mons et al. 2017) discuss, have begun to be adopted beyond science.

Wilkinson et. al. (Wilkinson et al. 2017) recently discussed a number of additions to FAIR including privacy protection via containers (based on Linked Data Platform containers)[3] and MetaRecords which return metadata given an HTTP GET request.

3 The FAIR TRADE Framework

Evaluation frameworks such as FAIR apply to both centralised and decentralised data solutions with little modification, but, in the decentralised case, we argue that they are insufficient. There are aspects of decentralisation which can have a significant effect on possible use of data, and which should be accounted for in any evaluation framework. To address this, we propose the FAIR TRADE framework for assessing decentralised data solutions. FAIR TRADE is an extension of FAIR: findability, accessibility, interoperability and reusability remain relevant dimensions for evaluation. In addition, we add dimensions of being TRusted, Autonomous, and Distributed and dEcentralised to form TRADE. Assessing a data solution according to FAIR TRADE provides a clear characterisation of its properties in terms of decentralisation and reusability.

While there is no essential need to modify the existing FAIR principles to apply to decentralised data, there are particular aspects worth noting that may require attention in interpreting them. For example, appropriate adaptations to ensure global uniqueness and persistence of identifiers may need to be made in a decentralised scenario to support Findability. The authorisation and authentication are harder in decentralized autonomous contexts, so they need more attention. The principles of Interoperability and Reusability are more straightforwardly agnostic of (de)centralisation status.

Our contribution

Decentralisation brings with it a number of further dimensions along which data management can vary, and which relate to the quality of a data source. We propose that the dimensions of trust, autonomy, distribution and decentralisation are effective aspects of data management to consider in decentralised contexts, and describing where in a FAIR TRADE space a data source lies is a useful characterisation of its nature, with adoption of *all* of the principles an indicator of best-practice in decentralised data stewardship.

1. ***TRusted*** – *T1: Data publication metadata can be automatically verified and validated* (e.g., that the publishing individual or organization, and timestamp of publication can be checked).

T2: *Claims made on the basis of data contents can be automatically verified* (e.g., that data describes the award of an educational certificate to the data owner). In interpreting this principle in specific contexts, the concepts and mechanisms of Verifiable Claims (as per the W3C Working Group) are useful: trustable access to properties of data, which is privacy-respecting, cryptographically secure and automatically verifiable.

[1] https://www.go-fair.org

[2] https://www.force11.org/community/members-directory

[3] https://www.w3.org/TR/ldp/#ldpc-container

Both of these principles are aided by support for a robust notion of identity, with attestations. This is important for data attribution and claim interpretation, and the various Verifiable Claim roles of holder, issuer, inspector-verifier and identity registry can all be relevant.

2. **Autonomous** – AU1: The owner(s), or authorized controller(s), of a piece of data have control over access to, and use of, data.

AU2: Data relating to personal identity follows AU1 and is trusted in the sense of T2.

Note that autonomy of data is not the same as accessibility under the FAIR principles. One may have access to, and have the credentials to have access to, one's data without having autonomous control over it. For example, personal data stored on the Facebook platform can be accessed by authorised users, but those users, including the person whose data it is, are not in control of it. In general, the principle of Autonomy relates to the legal and moral rights of data subjects over their own data.

3. **Distributed** – D1: In a given network, data is physically stored across some proportion or selection of nodes in that network.

D2: In a given network, the evaluation of data processing tasks can be executed across some specifiable proportion or selection of nodes.

Distribution is not necessarily a simple yes/no question: there are different degrees to which data, and data processing, can be distributed. Analysis of a system according to this principle supports, as appropriate, the characterization of quantified degrees of distribution for both D1 and D2 instead of a binary choice.

4. **dEcentralised** – E1: In a given network, no single node or small set of nodes controls which subset of data contents are held by any node or small set of nodes other than itself.

E2: In a given network, no single node or small set of nodes decides which new nodes can join.

A range of scenarios implementing the Trusted, Distributed and Decentralised principles are outlined in (Third and Domingue 2017), which we have since developed further into the LinkChains approach described in section 5.

4 Examples of Decentralisation

4.1 Distributed Ledges and Distributed Data Storage

It is important to distinguish between the terms 'distributed ledgers' and 'blockchains', which are often incorrectly used as synonyms. Distributed ledgers are replicated, shared and synchronised digital data geographically dispersed over multiple sites possibly including multiple institutions. A peer-to-peer network is required for communication and a consensus algorithm ensures replication and synchronisation across the multiple nodes. A key difference between applications that run on standard platforms and those that run on top of distributed ledgers is the way that data is stored and managed. Rather than connecting from a device (e.g. a mobile phone) to a central server, which holds all the required data (including private data), every player or volunteer in the network participates in ensuring that the whole network contains multiple copies of all the data. This changes a fundamental dynamic. The notion of centralised control disappears completely, rather data and computation are evenly owned, controlled and shared across the peer network. A blockchain is a specific type of distributed ledger where an ever-growing list of records, called blocks, are linked together to form a chain – hence the term 'blockchain'. The first blockchain was conceived by Satoshi Nakamoto in his white paper (Nakamoto 2008) as the basis for Bitcoin, the most famous blockchain based crypto-currency. The main idea behind Bitcoin was to create a currency specifically for the Internet rather than (as is the case in all fiat currencies) mapping an originally physical currency to the global communications infrastructure. Research produced by the University of Cambridge estimates that in 2017, there were 2.9 to 5.8 million unique users using a cryptocurrency wallet, most of them using Bitcoin.[4] The first issue that arises with Internet-based currencies is what is called the 'double spending problem'. This is the case when a digital 'coin' is spent, by an individual, for some service or good, and then the same coin is spent again by the same individual, for example, by copying or duplicating

[4] https://www.jbs.cam.ac.uk/faculty-research/centres/alternative- finance/publications/global-cryptocurrency

the relevant data. Blockchains address this problem by providing an immutable public ledger of all historical transactions. Once processed and stored within a block, and sufficient subsequent blocks have been confirmed, a transaction cannot be altered even by the transaction owners. Figure 1 shows a blockchain containing three blocks. Starting from the right, the newest block, each block points to its predecessor using a hash function. Transactions are stored in a Merkle Tree (Merkle 1980) - a tree of hashes - where the leaf nodes contain the transactions. This structure allows for efficient retrieval and ensures the veracity of the individual transactions in addition to the block – if a transaction is altered, then the hash will no longer be valid.

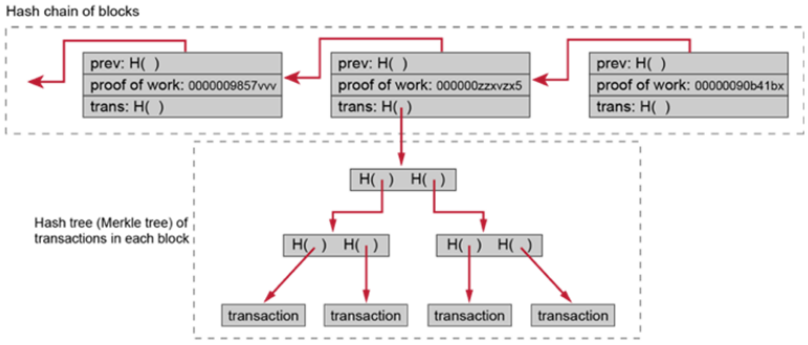


Figure 1: The hash links, proof-of-work nonce and Merkle tree of transaction data within a blockchain

The proof of work consensus mechanism which involves solving a cryptographic puzzle before anyone else has led to the growth of the computing power and electrical consumption associated with blockchain networks. Estimates are that by 2020 the Bitcoin network will expend as much electricity as Denmark.[5]b This has led to several blockchain platforms exploring other consensus mechanisms such as:

- ***Proof of stake*** – where the chances of being selected to produce the next block depend on the value of a 'stake' stored by a miner in a specific location. Variants of this take into account the 'age' of the stake.
- ***Proof of capacity*** – rather than the chances of being selected being related to the amount of computing power, as for proof of work, here the probability is related to the amount of storage a miner holds.
- ***Proof of burn*** – sending coins to an irretrievable address ('burn') gives one the right to be selected. The chances of being selected to mine the next block are related to the value of the burn.
- ***Proof of elapsed time*** – Intel has produced a special processor capability to implement a mechanism which relates elapsed time to the probability of being selected.

After Bitcoin, Ethereum (Buterin 2013) is the best known blockchain platform. Rather than serving as a platform for a cryptocurrency the underlying aim for Ethereum is to be an open blockchain platform to support the development and use of decentralised applications. Unlike Bitcoin, the programming language available on the Ethereum platform is Turing complete, so that general applications can be run on what the founders call a 'world computer'.

At the core of the Ethereum concept are two types of accounts:

- ***Externally Owned Accounts (EOAs)*** which are controlled by private keys. A private key is a cryptographic mechanism allowing for individuals to sign a transaction which has been secured by a corresponding public key. EOAs are controlled by individual users or organisations.
- ***Contract Accounts,*** also termed 'Smart Contracts' are controlled by contract code and are activated by EOAs.

[5] https://motherboard.vice.com/en_us/article/aek3za/bitcoin-could-consume-as-much-electricity-as-denmark-by-2020

When *ether*, the currency used within Ethereum, is sent from an EOA to a Contract Account, the contained program is executed. This can result in further transactions and payments and additional Smart Contracts being invoked. Through these chains of invocation, connected Smart Contracts form the basis of Ethereum applications called 'dApps' (distributed applications). A number of high-level languages exist for Smart Contracts including Solidity (similar to C/JavaScript), Serpent (similar to Python) and LLL (a low-level Lisp-like language).

One Smart Contract that has attracted attention recently is ERC721 ("Ethereum Request for Comments" #721).[6] This Smart Contract was developed after the Crypto Kitties[7] game became very popular. Players of Crypto Kitties can collect, trade and breed cats, with all activity running on the Ethereum blockchain. At its height, the platform raised $12M of venture capital and the most popular cats were valued at over $100,000. ERC721 extends ERC 20, a Smart Contract for representing tradeable tokens (such as coins). The key contribution of ERC721 is to represent *non-fungible* tokens - tokens which are each unique and therefore not interchangeable. Below we outline how we have extended this Smart Contract in our own work.

The Interplanetary File System (IPFS)[8] is a peer-to-peer distributed file system which is in some ways analogous to the Web, but which uses BitTorrent techniques for exchanging data. Each stored file is indexed by its hash with the indexes stored on network nodes. Human readable decentralised file naming is supported through IPNS analogous to DNS. The benefits brought by IPFS include content-based addressing, increased speed of data delivery over networks since large files can be transported in parallel, easy replication of valuable data, and network resilience in areas of low connectivity. IPFS has often been used together with blockchains as a data storage area and one recent joint venture is Filecoin[9] whereby peer data hosters are paid.

4.2 Verified Claims and Self-sovereign Identity

The Verifiable Claims Working Group (VCWG) aims to make the verifiable exchange of claims easier and more secure on the Web – in particular, how to make claims cryptographically secure, privacy respecting, and automatically verifiable. The main roles in the use of claims and their relationship are outlined in Figure 2, taken from (Sporny and Longley 2017)):

- ***Holder*** - the owner or controller of a number of verifiable claims, for example a student with qualifications or a citizen with personal attributes (e.g. age) or rights (e.g. to work).
- ***Issuer*** - creates verifiable claims, each connected to a specific subject, and then transmitted to claim holders. Examples here include universities (for qualifications) and governments.
- ***Inspector*** verifier - processes all received verifiable claims. Examples include employers and staff managing national borders.

Identifier registry - mediates the creation and verification of entities about which claims are made. Examples may include university records, organisational employee databases and governmental citizenship data stores. It is acknowledged that distributed ledgers can play the role of identifier registries in decentralised contexts. Standard identity management systems are based on centralised authorities such as certificate authorities or domain name registries. These are in some cases perceived as contributing to some of the problems outlined above, that is, contributing to the over centralisation of data. A partial community response to this has centred around the notion of 'self-sovereign identity' whereby users own, control, and manage their data. The main mechanism for this is through the following two main steps. Firstly, using a special app - an identity wallet - the user generates a public/private key pair which is unique and stored locally. Secondly, users request attestations using their public key from authorities and these are digitally

[6] http://erc721.org/

[7] https://www.cryptokitties.co/

[8] https://ipfs.io/

[9] https://filecoin.io/

signed by and stored on the wallet as well. An identity check from an identity authority may be conducted using mobile phone and passport to verify a user's identity. A blockchain can serve as a public trusted store for public keys. The main benefits associated with self-sovereign identity are that users maintain ownership and control over their data. Users can share attestations which can be shaped to minimise the exposure of private data. For example, a citizen may have an attestation, from an appropriate governmental agency, that they are 'over the legal drinking age' independent of any stored record of birth date.

Building on distributed ledger technologies, such as blockchains, the W3C Decentralised Identifiers Community Group have recently generated a data model and syntax for Decentralised Identifiers (DIDs) (Reed et al. 2018). DIDs are a new type of identifier for verifiable self-sovereign digital identity under the control of a DID subject, independent from any centralized authority. Each DID Document is comprised of three parts. *Cryptographic material* enables documents to be authenticated through a suite of *authentication suites. Service endpoints* support the provision of decentralised identity management services using the data contained in the DID.

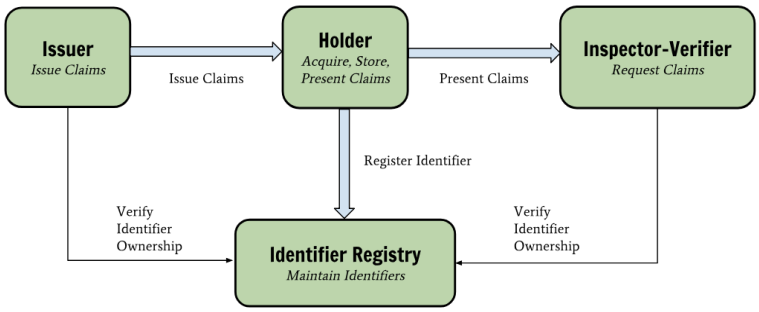


Figure 2: An example type of scenario for the use of verifiable claims. Taken from (Sporny and Longley 2017)

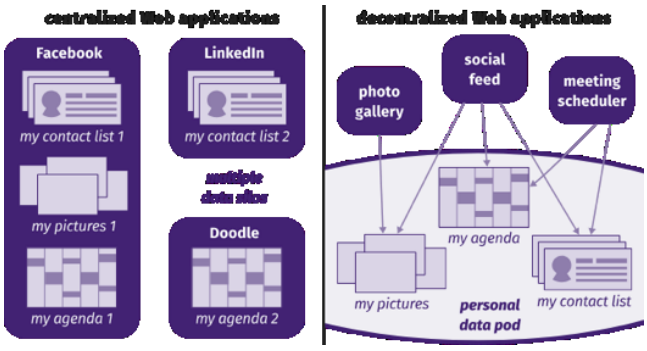


Figure 3: An overview of Solid concepts (taken from (Verborgh 2017))

4.3 Decentralised Linked Data Approaches

The Linked Data model supports, by design, the capacity for distributed or decentralised data systems. A single SPARQL query may, through the SERVICE keyword, pull data from multiple independent sources and aggregate it to produce a single set of results. Linked Data approaches in general, however, are agnostic with regard to distribution and decentralisation and can be used to implement many variations along these spectra. In terms of approaches specifically designed with decentralization in mind, we discuss Solid and Linked Data Fragments.

Solid (Lee 2018b) is a new initiative from the inventor of the Web which aims to enable the decentralisation of personal data. As can be seen in Figure 3, each Solid user has their own personal data which is called a data pod. Users are free to choose who hosts the data - whether to host personally or with

a third party of choice. In the Solid scenario, applications are decoupled from the data they consume, interacting through APIs, with user permission, avoiding vendor lock-in and allowing users to easily switch between applications and servers without loss of service. An additional benefit is that newcomers can innovate and develop new applications for users to try out. As stressed in an open letter (Lee 2018c), Solid is driven by the principle of "personal empowerment through data".

According to Verborgh et al., (Verborgh et al. 2016) a Linked Data Fragment is a subset of an RDF based knowledge graph computed by some means. As stated by the authors the problem with most approaches to obtaining Linked Data Fragments outside of a centralized context is that the burden of computation rests entirely with the client, for example, if the data is downloaded, or with the server, for example, using SPARQL, as shown at the top of Figure 4. As shown at the bottom of Figure 4, Triple Pattern Fragments (Verborgh et al. 2016) allow the computation to be shared between a server and client. In essence the approach lowers the burden on RDF servers since the only computation required is on the simple matching of triples. The Triple Fragment Server returns matching triples in reasonably sized fragments (100 triples approximately) and, in addition, metadata and controls are returned. The metadata includes an estimate of the number of triples that match the given triple pattern. Controls enable clients to retrieve further triples from the same knowledge graph. A Linked Data Fragments client supports standard SPARQL queries over any returned triples. Triple Pattern Fragments make it easy to query across multiple data stores at once - much as with SPARQL's SERVICE keyword, but with the bulk of the computational work of aggregation moved from server to client. Triple Pattern Fragments offer a specific advantage for decentralised querying compared to SPARQL, in that it is no longer *required* to know in advance which pieces of data are to be found on which federated endpoint.

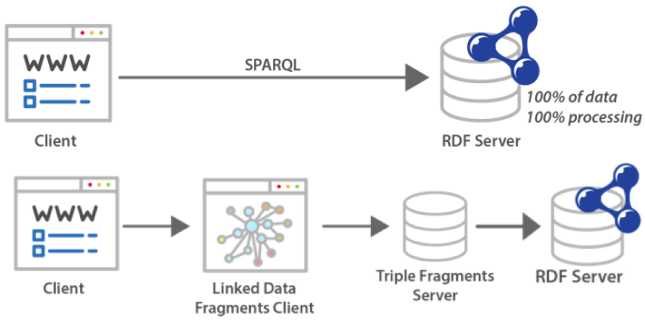


Figure 4: At the top, a standard SPARQL scenario where an RDF server is responsible for hosting and processing all the data. At the bottom, a combination of a Linked Data Fragments client and a Triple Fragments server allows the burden of computation to be shared between a client and server.

4.4 LinkChains

In designing LinkChains, our overall goal has been to bring together the benefits accrued from LD approaches together with decentralised approaches, as embodied by distributed ledgers and IPFS. As can be seen in Figure 5, the LinkChain architecture has three main parts: a user-controlled private storage area; a decentralised public and private data storage area; and a blockchain.

We currently use Solid personal data pods to store private data solely under user control (complete control over the pod-stored data), although LinkChains will talk to any Linked Data Platform.[10] Read/write access to Solid pods is granted through the Solid auth client (Lee 2018b). Non-private or encrypted data is stored within a decentralised data storage system, as opposed to directly on a blockchain, as most public blockchains require payment for data storage and do not store large volumes of data efficiently. We use a blockchain instead to record and verify claimed data attributes, and to facilitate the transfer of data as an asset.

[10] https://www.w3.org/TR/ldp/

A LinkChain Smart Contract enables data on the blockchain to be written or queried. The key functionality provided through our Smart Contract is the ability to store LinkChain Trust Tokens (LCTT) - tokens verifying a set of attributes of a Linked Data set. In particular, a LCTT provides signed verification that declared attributes hold for a Linked Data set. LCTT are created by extending the ERC 721 token standard described above. The main parts of this LinkChain architecture are controlled through a Web/mobile application or a client using the APIs and interfaces provided by the LinkChain platform.

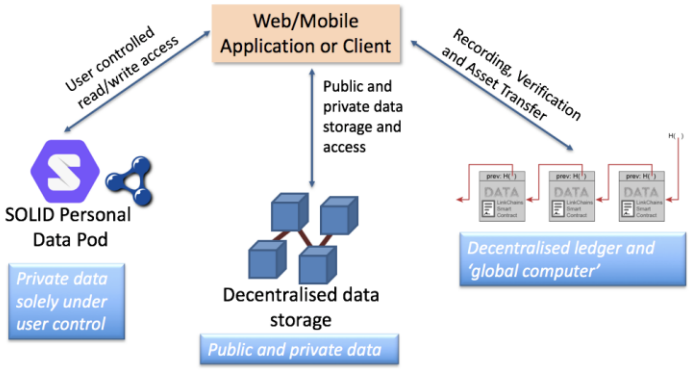


Figure 5: The overall LinkChain architecture which comprises of three main parts: storage of private data under user control; decentralised public and encrypted private data; and a blockchain for recording, verification and asset transfer.

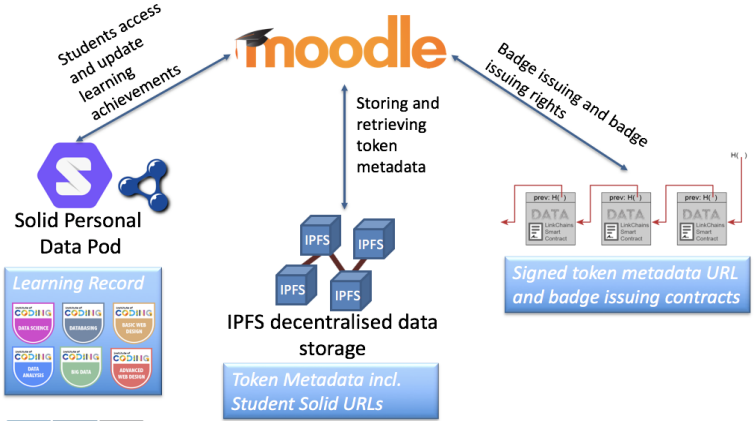


Figure 6: An instantiated LinkChain architecture supporting the storage of personal learning records including Open Badges.

Figure 6 shows how we instantiated the above generic LinkChain architecture to store and verify lifelong learning records. This ongoing work is a part the Institute of Coding[11], an initiative from the UK government to tackle the UK digital skills crisis.[12] When a student completes a particular assignment using the Open University's Virtual Learning Environment (VLE), an Open Badge (based on the IMS Open Badge standard)[13] is automatically issued which students can consequently claim and have stored within their Solid data pod. Additionally, a digitally signed hash of the badge is encoded within an LCTT placed

[11] https://instituteofcoding.org/

[11] https://londonlovesbusiness.com/institute-of-coding-launches-to-tackle-uk-digital-skills-crisis/

[13] https://www.imsglobal.org/cc/statuschart/openbadges

onto the LinkChain blockchain and a token metadata file is stored on IPFS. Placing the metadata on IPFS rather than onto the blockchain reduces cost and increases the overall data handling efficiency. Figure 7 shows a screen snapshot of the student interface used when the system was deployed in our 2018 Summer of Code online school.[14] We can see in Figure 7 that the student has selected her Super Badge, given for successfully completing all nine assignments, and can view the details of the blockchain badge representation. Figure 7 also shows that each element of the Open Badge, including *assertion*, *recipient* and *evidence*, has its own distinct blockchain representation encoding data such as the date of issue and whether the badge has been revoked or has expired.

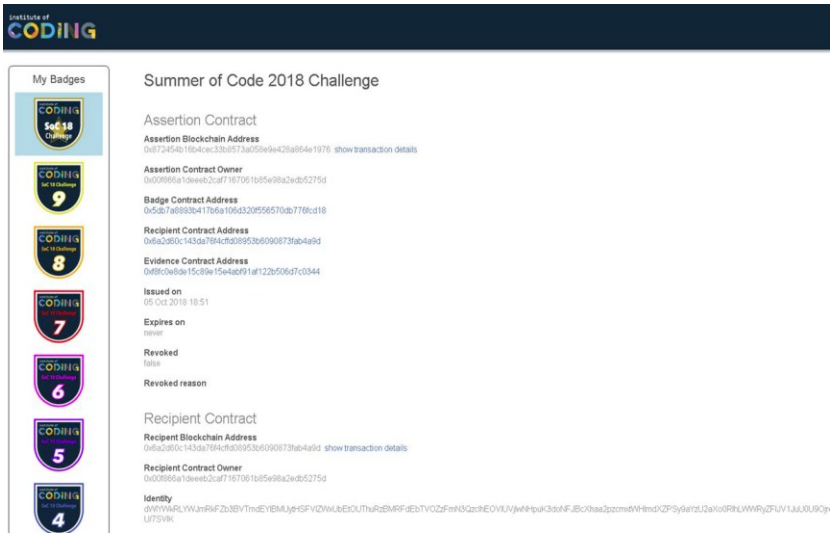


Figure 7: A portion of the interface for OU students viewing their blockchain badges on the 2018 Summer of Code course.

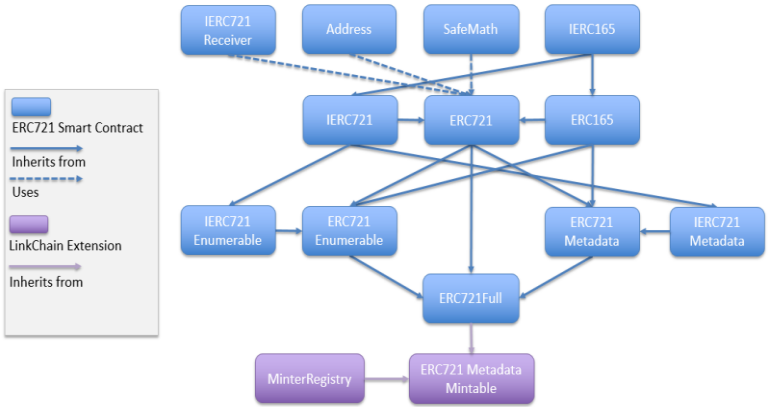


Figure 8: Our extensions to the ERC 721 Smart Contract to support our LinkChains implementation

Figure 8 shows ERC721, including the extension we created for issuing the non-fungible tokens within our LinkChain platform. Each rectangle in the image represents a distinct Smart Contract and the set of blue rectangles together form the ERC721 Smart Contract Template labelled ERC721Full. This template is part of OpenZeppelin[15], a framework of reusable smart contracts for Ethereum and Ethereum Virtual Machine

[13] http://www.open.ac.uk/about/teaching-and-learning/esteem/projects/themes/supporting-students/summer-code

[14] https://github.com/OpenZeppelin/openzeppelin-solidity/blob/master/contracts/token/ERC721/ERC721.sol

(EVM) based blockchains (i.e. blockchains which use the EVM as an execution platform). ERC-721 defines a minimum interface a smart contract must implement to allow unique tokens to be managed, owned, and traded. IERC721 is an interface which defines the functions used in the smart contracts and is used as a way to implement those functions. The "address" smart contract is used to separate account addresses and smart contract addresses. The SafeMath smart contract makes sure that the mathematical operations performed in a smart contract are safety checked which can revert on error. ERC165 standard interface detection is used to expose the interfaces that an ERC721 smart contract supports. The ERC721 Enumerable extension provides functionalities to sort through the tokens easily, whereas the Metadata extension enables a token to have a contract name, symbol and some extra data that makes it unique.
The two smart contracts MinterRegistry and ERC721MetadataMintable (highlighted in purple) together form the core of our LCCT token implementation. MinterRegistry contains a list of addresses which are permitted to mint, i.e. are allowed to issue new ERC721 tokens. ERC721MetadataMintable issues LCCT tokens upon receiving legitimate minter requests in the form of metadata representing the address of the minter, the address that the tokens created are to be sent to, the address of the token contract and the token ID. On ERC721MetadataMintable being invoked, the smart contract authenticates the minting address and issues the LCCT token to the specified address.

5 Evaluation according to FAIR TRADE

We evaluate these principles by considering a number of decentralised data systems and identifying which of the principles, and subprinciples, apply in each case, and to what degree. The goal is to test their *expressive power*: in applying these principles, can we describe significant differences with regard to various systems and approaches? The choice of systems to evaluate is diverse, including systems with significantly different goals and designs, deliberately to illustrate this expressive power.

5.1 Solid

Solid accrues many of its positive scores with respect to the FAIR TRADE framework due to the fact it is a set of conventions and tools based on Linked Data principles and more generally as far as possible on existing W3C standards and protocols. Specifically, the use of URIs supports *findability* through the provision of a unique and persistent identifier. *Accessibility* is supported through the return of RDF based metadata, understandable to both humans and machines, available through HTTP requests and SPARQL queries. Metadata in a formal knowledge representation language such as RDF also supports *interoperability*. The PROV-O provenance vocabulary[16] and the Linked Data Rights vocabulary[17] can support *re-usability*.
We take the relevant network, in the case of Solid, to be the collection of all data pods. *Distribution* and *decentralisation* of data are therefore satisfied for all subprinciples. Each is a distinct per-user Web server allowing users in principle to be completely *autonomous* in terms of data control (AU1), but without a stronger model for identity, AU2 does not hold. Solid does not satisfy any aspects of the *trusted* dimension. This highlights the fact that there is no inbuilt mechanism for verifying that any claims made within a data pod are true. The Verifiable Claims data model can be used to express claims, but additional infrastructure would be required to support verification.

5.2 DBpedia Infrasructure

DBpedia (Lehmann et al. 2015) is, at its core, a machine-readable formal representation of the various Wikimedia projects including Wikipedia. Data within DBpedia is served as Linked Data, enabling semantic query processing. The English version of the DBpedia knowledge base contains over 4.5 million objects, over 4.2 million of which are classified within the DBpedia ontology. DBpedia is not, in itself, of course, a

[16] https://www.w3.org/TR/prov-o/

[17] http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/

data platform; it is a dataset. Nonetheless, there is an infrastructure surrounding its creation, management, and publication, including a central DBpedia site[18] and a network of mirrors. We discuss this infrastructure here.

As it is founded on Linked Data, DBpedia is positively assessed against all the FAIR criteria. Every DBpedia data item has a unique dereferencable URI which contains data and metadata represented in RDF satisfying the *findability*, *accessibility* and *interoperability* dimensions. All data in DBpedia is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License supporting the *re-usability* principle.

Trust is generated from the fact that the data has been extracted from Wikipedia, and that Wikipedia generates trust for the information it holds from its inbuilt consensual and social mechanisms.[19] However, the "automatic" aspects of T1 and T2 are not satisfied. DBpedia data is centralised, and ultimately controlled by the DBpedia Board of Trustees,[20] which currently has eight members; data is not *autonomous* with respect to contributors or users in the senses of AU1 and AU2. DBpedia provides tools to make it easy to maintain a live mirror,[21] however, synchronisation is one-way (central data is pushed to the mirrors); D1 and E2 are satisfied, but D2 and E1 are not.

5.3 Linked Data Fragments

As with the systems described above, Linked Data Fragments passes all the FAIR criteria as it is based on Linked Data, but only partially satisfies the new criteria. As with Solid, there are no inbuilt mechanisms to facilitate *trust* of any provenance for (T1), or claims made about (T2), returned data. Following the core aim of Linked Data Fragments, the platform facilitates the distribution of data between clients and servers - primarily to enable the sharing of computation association with processing. Linked Data Fragments certainly supports both D1 and D2 with regard to *distribution*, and E1 and E2 with regard to *decentralisation*. With regard to *autonomy*, data on a Linked Data Fragments server is controlled by whomever controls the server. and there is no standard means of controlling data visibility. AU1 does not, therefore, hold, nor does AU2: there is no consideration of user identity in the standard at all.

5.4 Bitcoin

Bitcoin core can be used to find transaction/exchange data. But, the metadata or message embedded in the OP_RETURN part of the bitcoin transaction does not lend itself to findable data. Bitcoin as a platform contains no internal indexing or search mechanisms - all discovery is down to external tools which are fixed, or domain specific, or proprietary (e.g. Blockchain.com) [22]; thus transaction/exchange data is *findable*, but data embedded within a Bitcoin transaction is not. As Bitcoin is a public blockchain, all data is *accessible* from the Bitcoin blockchain at any time. No format is prescribed for representing embedded data on Bitcoin, nor is there support for vocabularies; embedded data in a Bitcoin transaction is in general is not *interoperable*. The fact that every Bitcoin transaction including the embedding of data is signed by a private key means that data has at least some verified provenance. This makes embedded data in a Bitcoin transaction *re-usable* to some extent, although it has no standard license, or licensing infrastructure.

Bitcoin miners are distributed over the world and every node has a copy of the entire blockchain. The combination of peer data replication, the proof of work consensus mechanism, and immutability means that Bitcoin data can be *trusted* – neither provenance metadata (T1) nor data contents (T2) can be modified once published. *Autonomous* control of data directly on-chain is limited; data cannot be deleted from the

[18] http://dbpedia.org

[19] https://en.wikipedia.org/wiki/Wikipedia:Consensus

[20] https://wiki.dbpedia.org/board

[21] https://github.com/dbpedia/dbpedia-live-mirror

[22] https://www.blockchain.com/explorer

blockchain, although some control over visibility can be implemented using encryption. Hashes with pointers to off-chain data offer greater potential for autonomy, as does the use of multiple anonymous accounts via a management wallet, but these, particularly the former, require infrastructure going beyond Bitcoin itself. We cannot therefore say that it satisfies either of the autonomy subprinciples The Bitcoin ledger is globally *distributed*: every node has a copy of the full blockchain (D1) and every node executes every script (D2), and fully *decentralised* (E1 and E2). There have been concerns within the Bitcoin community that an individual or consortium of miners could take over the network by gaining more than 50% of the total hashing power and thereby the ability to rewrite history. Thus far, this has not happened; at the time of writing, all mining groups have less than 20% control of the overall network.[23]

5.5 Ethereum

As Ethereum and Bitcoin share similar features, its evaluation results according to FAIR TRADE are similar to Bitcoin. A noteworthy difference is with regard to *autonomy*: the availability of smart contracts gives Ethereum increased power in terms of what the blockchain network itself can support; the autonomy subprinciples themselves may potentially be easier to support on Ethereum with a self-sovereign identity and Verifiable Claims infrastructure using smart contracts. Nonetheless, further work, and off-chain infrastructure, would be required in order to support both AU1 and AU2 fully.

5.6 LinkChains

LinkChains is designed to combine the user-focused Linked Data design of Solid with a blockchain-backed trust layer, and thus inherits traits from both Solid and Ethereum. That it meets more of the FAIR TRADE criteria than other systems here is not a surprise; a FAIR TRADE analysis played a role in the development process. Rather, it provides an example of how this framework can focus efforts by making explicit the principles for high-quality decentralised data management. The key element *not* yet present in the LinkChains design for full FAIR TRADE compliance is support for *autonomous* identity, although we are currently exploring solutions for self-sovereign identity approaches to address this gap.

Table 1: Evaluation according to FAIR TRADE

	Solid	DBpedia	LinkChains	Bitcoin	Ethereum	Linked Data Fragments
Findable	Via LD	Yes	Via Solid	No internal index or search	Similar to Bitcoin.	Yes
Accessible	Via LD	Yes	Via Solid	Yes	Yes	Yes
Interoperable	Via LD	Yes	Via Solid	Data in arbitrary format	Data in arbitrary format	Yes
Re-usable	Via LD	Yes	Via Solid	No licensing.	No licensing.	Yes

[23] https://www.blockchain.com/en/pools

Trusted	Neither T1 nor T2.	Neither T1 nor T2; not automatic.	T1 and T2, via blockchain.	T1 and T2, by blockchain design.	T1 and T2, by blockchain design.	Neither T1 nor T2.
Autonomous	AU1, not AU2.	Neither AU1 nor AU2.	AU1 via Solid, not AU2.	Neither AU1 nor AU2.	Neither AU1 nor AU2.	Neither AU1 nor AU2.
Distributed	D1 and D2.	D1, not D2	D1 and D2.	D1 and D2	D1 and D2.	D1 and D2
Decentralised	E1 and E2.	Not E1. E2.	E1 and E2	E1 and E2.	E1 and E2.	E1 and E2.

6 Conclusions

As we stated at the start of this paper, a growing number of societal concerns have been raised over the last few years related to how the centralisation of data has led to the loss of user control and the dangers associated with data misuse. A number of different technical communities associated with peer-to-peer and Web technologies have been responding to this in part with a variety of new approaches and platforms to decentralising data. The contribution of this paper is to provide a framework to begin to compare systems, based on an explicit set of principles relating to decentralised data management.

Above, we have assessed a range of data solutions for adherence to the FAIR TRADE principles. As well as a measure of quality in data stewardship, the exercise of assessment showed that these principles, even when not fully followed, provide a good characterisation of the properties of the data or platform being assessed. One can see at a glance, for example, that the DBpedia project is concerned with providing a common resource -- concerns relating to autonomy do not apply when the goal is to elicit effectively *donations* of information for public use, but those relating to trust, openness and reuse do apply, with distribution only as required to ensure the accessibility of data. The Bitcoin and Ethereum platforms meet most of the TRADE standards, except that of being Autonomous, but not FAIR, reflecting their development from cryptocurrencies and, in Ethereum's case, distributed computation - their use as decentralised data platforms coming later. Solid and Linked Data Fragments are evaluated to be very similar, reflecting their shared concerns and foundations on Web data standards. It is perhaps unsurprising to see that the platforms which most fit the FAIR TRADE principles are those stemming from a field in which the focus is on data publication and standards. LinkChains, although a work in progress, is being designed to be FAIR TRADE from the outset, and is based on a combination of other technologies, each of which follows some of the principles. In particular, combining Solid's focus on individual data with the possibilities for trusted and verifiable data based on blockchains, distributed identity and Verifiable Claims has the potential to provide a flexible platform with best-practice principles in data management in the decentralised sphere. Note that these systems did not stretch the descriptive capacity of the FAIR TRADE principles: as noted earlier, each of the presented principles can cover a range of possible useful or interesting behaviours which a distributed data solution could display. As novel technologies arise, we expect this framework to continue to provide a useful common vocabulary for assessing and comparing decentralised data systems. By taking a forward-looking approach to data management standards in a decentralised context, we aim to contribute to the establishment of best practice and well-motivated technical approaches early in the development of the field, at a time when their adoption is significantly easier than in a more entrenched technical landscape.

REFERENCES

Asano, Y., Hidaka, S., Hu, Z., Ishihara, Y., Kato, H., Ko, H.-S., Nakano, K., Onizuka, M., Sasaki, Y., and Shimizu, T., 2018. A View-based Programmable Architecture for Controlling and Integrating Decentralized Data. *arXiv preprint arXiv:1803.06674*.

Bureau, C. F. P., 2015. *Data Point: Credit Invisibles* [online]. Available from: https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

Buterin, V., 2013. *Ethereum White Paper* [online]. Available from: https://github.com/ethereum/wiki/wiki/White-Paper [Accessed 10 Dec 2018].

Curran, D., 2018. Are you ready? Here is all the data Facebook and Google have on you. *The Guardian* [online], 2018. Available from: https://www.theguardian.com/commentisfree/2018/mar/28/all-the-data-facebook-google-has-on-you-privacy [Accessed 10 Dec 2018].

Fact, F., 2018. Windrush generation: what's the situation? [online]. Available from: https://fullfact.org/immigration/windrush-generation/ [Accessed 10 Dec 2018].

Grierson, J., 2018. Windrush row: 63 people could have been wrongly removed. *The Guardian* [online], 2018. Available from: https://www.theguardian.com/uk-news/2018/may/15/windrush-row-63-people-could-have-been-wrongly-removed-says-javid [Accessed 10 Dec 2018].

Guardian, T., 2015. The Cambridge Analytica Files. *The Guardian* [online], 2015. Available from: https://www.theguardian.com/news/series/cambridge-analytica-files [Accessed 10 Dec 2018].

Lee, T. B., 2018a. Interview in Vanity Fair. [online]. Available from: https://www.vanityfair.com/news/2018/07/the-man-who-created-the-world-wide-web-has-some-regrets [Accessed 10 Dec 2018].

Lee, T. B., 2018b. SOLID Project Website. [online]. Available from: https://solid.mit.edu/ [Accessed 10 Dec 2018].

Lee, T. B., 2018c. One Small Step for the Web. [online]. Available from: https://www.inrupt.com/blog/one-small-step-for-the-web.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., and Auer, S., 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2), 167–195.

Merkle, R. C., 1980. Protocols for public key cryptosystems. *In*: *Security and Privacy, 1980 IEEE Symposium on*. IEEE, 122.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., and Wilkinson, M. D., 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37 (1), 49–56.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system.

Özsu, M. T. and Valduriez, P., 2011. *Principles of distributed database systems*. Springer Science & Business Media.

Reed, D., Sprony, M., Longley, D., Allen, C., Grant, R., and Sabadello, M., 2018. Decentralized Identifiers (DIDs) v0.11 Data Model and Syntaxes for Decentralized Identifiers (DIDs). *W3C* [online]. Available from: https://w3c-ccg.github.io/did-spec/ [Accessed 10 Dec 2018].

Schneider, F. B. and Trustworthiness, C. on I. S., 1999. *Trust in cyberspace*. National Academy Press Washington, DC.

Sporny, M. and Longley, D., 2017. *Verifiable Claims Data Model and Representations* [online]. Available from: https://www.w3.org/TR/2017/WD-verifiable-claims-data-model-20170803.

Third, A. and Domingue, J., 2017. LinkChains: Exploring the space of decentralised trustworthy Linked Data.

Verborgh, R., 2017. Paradigm shifts for the decentralized Web. [online]. Available from: https://ruben.verborgh.org/blog/2017/12/20/paradigm-shifts-for-the-decentralized-web/ [Accessed 10 Dec 2018].

Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., and Colpaert, P., 2016. Triple Pattern Fragments: a low-cost knowledge graph interface for the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37, 184–206.

Walport, M., 2015. *Distributed Ledger Technology: beyond block chain* [online]. Government Office for Science. Available from: https://pubs.acs.org/doi/pdf/10.1021/acsaem.8b00240%0Ahttp://pubs.acs.org/doi/10.1021/acsaem.8b00240.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., and Bourne, P. E., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Wilkinson, M. D., Verborgh, R., da Silva Santos, L. O. B., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., van Mulligen, E. M., and Ciccarese, P., 2017. Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, 3, e110.

Towards a Decentralized, Trusted, Intelligent and Linked Public Sector: A Report from the Greek Trenches

Themis Beris
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens
Athens, Greece
tberis@di.uoa.gr

Iosif Angelidis
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens
Athens, Greece
iosang@di.uoa.gr

Ilias Chalkidis
Athens University of Economics and Business
Athens, Greece
ihalk@di.uoa.gr

Charalampos Nikolaou
Dept. of Computer Science, University of Oxford
Oxford, United Kingdom
babis.nikolaou@cs.ox.ac.uk

Christos Papaloukas
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens
Athens, Greece
christospap@di.uoa.gr

Panagiotis Soursos
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens
Athens, Greece
psoursos@di.uoa.gr

Manolis Koubarakis
Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens
Athens, Greece
koubarak@di.uoa.gr

ABSTRACT

This paper is a progress report on our recent work on two applications that use Linked Data and Distributed Ledger technologies and aim to transform the Greek public sector into a decentralized, trusted, intelligent and linked organization. The first application is a re-engineering of Diavgeia, the Greek government portal for open and transparent public administration. The second application is Nomothesia, a new portal that we have built, which makes Greek legislation available on the Web as linked data to enable its effective use by citizens, legal professionals and software developers who would like to build new applications that utilize Greek legislation. The presented applications have been implemented without funding from any source and are available for free to any part of the Greek public sector that may want to use them. An important goal of this paper is to present the lessons learned from this effort.

CCS CONCEPTS

• **Information systems → Distributed storage**; **Semantic web description languages**; • **Security and privacy**; • **Computing methodologies → Knowledge representation and reasoning**; • **Applied computing → E-government**; • **Software and its engineering → Peer-to-peer architectures**;

KEYWORDS

linked open data; distributed ledgers; blockchain; open government; semantic web; bitcoin; tamper-proof; public sector services

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317077

ACM Reference Format:
Themis Beris, Iosif Angelidis, Ilias Chalkidis, Charalampos Nikolaou, Christos Papaloukas, Panagiotis Soursos, and Manolis Koubarakis. 2019. Towards a Decentralized, Trusted, Intelligent and Linked Public Sector: A Report from the Greek Trenches . In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317077

1 INTRODUCTION

Decisions made by governments and public authorities affect significantly the daily lives of ordinary citizens. Therefore, an important dimension of open and trusted government is making these decisions *open and easily accessible* to the public, so that citizens can trust the public organizations that make these decisions, and can easily check whether taxpayer's money is spent wisely.

We believe that Semantic Web, Linked Data and Distributed Ledger Technologies can help realize this vision of open and trusted government. However, the vision cannot be realized by these technologies alone: we also need other technologies (especially other Artificial Intelligence technologies) as we will discuss here.

Our experience in using ontologies, linked data and distributed ledger technologies in the public sector has been gained through project DiavgeiaRedefined[1], an internal project of our team in the National and Kapodistrian University of Athens.[2]

Diavgeia (https://diavgeia.gov.gr/en) [3] is a Greek program introduced in 2010, enforcing transparency over the government and public administrations, by requiring that all government institutions have to upload their decisions on the Diavgeia Web portal. The portal is managed by the Ministry of Administrative Reform. Diavgeia is now fully implemented by public authorities in Greece. The current rate of uploads in the portal is up to 42,000 decisions during busy working days, summing up to a total of 32.5 million up

[1] http://pyravlos-vm5.di.uoa.gr/diavgeia/
[2] http://kr.di.uoa.gr/
[3] διαύγεια is the word for transparency in Greek.

to now. These decisions come from more than 4,200 public sector organizations in Greece with over 91,000 accounts on the portal.

Decisions are currently uploaded in Diavgeia as PDF files and follow no structuring of their textual content. As a consequence, interested parties (the government, public authorities, ordinary citizens, non-government bodies, courts, the media, etc.) rely on keyword search over PDF files, in order to find decisions that might affect them in some way or verify that uploaded decisions have been taken according to the law. Also, despite the fact that these decisions are digitally signed, there is no integrity mechanism which ensures the immutability of all decisions over time.

Interaction with Diavgeia can take a lot of the precious time of public servants as our administrative colleagues at the National and Kapodistrian University of Athens have told us repeatedly. Combined with the bureaucratic apparatus of Greece, time spent on Diavgeia is, not only slowing down administrative processes, but also making the working life of public sector employees miserable. As an example, our colleagues estimate that getting authorization to travel to an international conference today at our university takes at least 15 days and involves at least 5 very experienced employees for at least 3 person hours.

So, does this means that Diavgeia needs to be abolished? We don't think so! Transparency is an essential element of public life and we need to fight for it as citizens and professionals. It is very important, especially in Greece, which has made a bad name for itself after the 2008 financial crisis, which resulted in the financial bailout of the country and the resulting unprecedented austerity imposed on the Greek people by the troika of lenders. We believe that disruptive technologies from Artificial Intelligence and Distributed Ledgers can actually help alleviate the inefficiencies, both technical and social, that are currently present in Diavgeia and lead to a state where Diavgeia fulfills its original purpose: to enable Greek taxpayer's money to be spent transparently, effectively and within the boundaries of the law.

With our work in DiavgeiaRedefined, we aim at revolutionizing the way that decisions of the Diavgeia program are made public, by following the footsteps of another successful research effort of our team, which we also discuss briefly in this paper. This effort has resulted in the development of the platform Nomothesia [4] that makes Greek legislation available on the Web as open linked data[5] [1, 6]. By applying Artificial Intelligence techniques, we envision a new state of affairs in which public sector employees and ordinary citizens have user-friendly speech- and natural language-based interfaces to government portals, such as Diavgeia, that can help them interact with public sector organizations in the most efficient and effective way. In addition, through the use of techniques from Distributed Ledgers, we enable public sector organizations to decentralize their decision making while at the same time guaranteeing unprecedented levels of trust and transparency. For example, in DiavgeiaRedefined, we use the bitcoin blockchain to enable Greek public sector decisions to remain immutable, introducing a great level of transparency to the Diavgeia program and ensuring the integrity of the published decisions as open linked data.

Technically, we view public sector decisions as a collection of legal documents which are organized according to an ontology and are encoded in RDF. As a result, decisions can be interlinked among themselves and with other open data in complex ways. For instance, a decision might refer to an earlier version of itself or it might refer to other legislative documents (such as laws or other public sector decisions) that are related to it. In addition, decisions might refer to other open linked data sources, e.g., the administrative divisions of Greece. Linking to other linked open data is crucial, because a software developer is then able to pose queries such as "Find all funds provided by the Greek government to municipalities with a population over 100,000 people" which targets data from the Diavgeia portal but also data from the Greek administrative geography dataset.[6] Our aim is to re-engineer the existing Diavgeia portal so that, not only it enables software developers to develop applications easily using data from Diavgeia, but it also gives ordinary citizens a way to verify public sector decisions are taken according to the Greek law, and that prudent administration and governance practices are followed. By adopting the tamper-proof nature of bitcoin in Diavgeia, we build tools that are able to commit Diavgeia's daily decisions in a single blockchain transaction and are also able to automatically validate the decisions' integrity.

The goal of this paper is to discuss the current state of play of our research efforts DiavgeiaRedefined and Nomothesia. We give more information about DiavgeiaRedefined since it fits perfectly with the theme of this workshop. We also discuss Nomothesia, since the two projects have contributed and continue to contribute to each other. In previous published work by our team, DiavgeiaRedefined has been presented in [2] and Nomothesia in [1, 6].

The rest of the paper is organized as follows. Section 2 discusses related works that combine linked data with blockchain technology. Section 3 presents the current Diavgeia portal and discusses its weaknesses. Section 4 discusses the DiavgeiaRedefined ontology, presents its web editor and visualizer modules and some interesting SPARQL queries. Section 5 describes the two blockchain tools developed for the preservation and verification of decisions. Section 6 presents our evaluation results that show that DiavgeiaRedefined can offer more functionality and efficiency than the current Diavgeia portal. Section 7 discusses the current state of play of DiaveiaRedefined's sister project Nomothesia. Section 8 discusses what other Artificial Intelligence technologies are required to realize the vision of intelligent, trusted, decentralized, linked public administrations and governments. Lastly, Section 9 summarizes the lessons we learned from DiavgeiaRedefined and Nomothesia and our efforts to make them available for free to the Greek public sector.

2 RELATED WORK

The Semantic Web and Linked Data research community has contributed a lot in the last few years in the domain of developing intelligent solutions for the public sector especially in Europe. Currently, the main such activity in Europe is ISA[2] which studies the interoperability of information systems deployed in the public sector in Europe.[7].

[4] Νομοθεσία is the word for legislation in Greek.
[5] http://legislation.di.uoa.gr/
[6] http://linkedopendata.gr/dataset/greek-administrative-geography
[7] https://ec.europa.eu/isa2/home_en

The Semantic Web and Linked Data community has also recently begun to consider applications which leverage the distributed, immutable nature of blockchains. In this context, one workshop examining the interaction of these two areas took place in the Web Conference of 2018 in Lyon[8] and it is continued in 2019 by the present workshop.

The article [9] was the first to discuss what Semantic Web research and development can offer to Blockchain research and development and vice versa. The paper [10] examines a semantic approach for decentralized semantic identities (in the context of the W3C WebID effort), in which the Namecoin blockchain is used to register the user's WebID URI and domain names. In this endeavor, the proposed authentication scheme is outside the control of any single entity. The paper [22] proposed a linked-data-based method of utilizing blockchain technology to create tamper-proof audit logs that provide proof-of-log manipulation and non-repudiation.

In [18], Janowicz et al. discuss the value proposition of distributed ledger technologies and crypto-currencies for academic publishing. An informal model of how the Semantic Web journalâĂŹs peer-review workflow could benefit from distributed ledger technologies is provided, while the challenges of implementing a system of this kind are documented. While the problem addressed here is oriented towards academics, the authors argue that the proposed methodology can be applied to any similar science-related challenge.

Finally, Hoffman et al. tackle the challenge of implementing a permissioned blockchain-based system[17] aimed at eliminating losses that tax authorities are currently struggling with. A key innovation of this work is the proposal of a VAT Invoice 2.0 modelled as a Linked Data document. The main idea is that, having an ontology-based tax document and the right permissions, one can investigate the entire commercial chain for any taxable item.

3 BACKGROUND ON DIAVGEIA

In this section, we present the current Diavgeia portal in detail and point out the problems of the current implementation.

3.1 Greek public sector decisions

Public sector decisions cover a broad spectrum of activities in Greece. The Greek government has enacted 34 different decision types that may be uploaded on Diavgeia. The decision type is chosen by the government institutions according to the context of the decision. Despite the many different decision types, we observed that the majority of them follow the same pattern. A decision starts by refering to a number of different Greek laws on which is based[9] and then gives the main text of the decision. The following figure illustrates an example of an *Appointment* decision type that adheres to the aforementioned pattern.

Example of an *Appointment* decision type

Appointment of R.F. as Full Professor

In accordance with:

(1) The provisions of Law 3549/2007, article 25, paragraph 1.
(2) The provisions of Presidential Decree 2011/54.
(3) The provisions of Law 4386/2016, article 70, paragraph 4.

We decide:

(1) The appointment of R.F. as Full Professor at the X department, at the Y University, on the subject of "Semantic Web".

Despite the fact that this pattern can be used to define a common format for the different types of decisions, for the time being, public sector authorities upload their decisions as PDF files which follow no structuring of their textual content. Futhermore, citizens have no guarantee that the legislative references of a decision exist and are valid (such as *Laws* and *Presidential Decree* of the appointment example). By using the 5-star rating model for data [3], Greek public sector decisions are marked as 1-star.

In this work, we improve the current way of publishing Greek public sector decisions on the Web, by expressing them as 5-star open linked data. We use the aforementioned pattern as a basis to develop the Diavgeia ontology. Technically, we view decisions as a collection of legal RDF documents with this standard structure.

We also employ the Nomothesia in order to ensure that the references to Greek legislation exist. Nomothesia has so far published 5 primary types of Greek legislation (*Constitution*, *Presidential Decrees*, *Laws*, *Acts of Ministerial Cabinet*, and *Ministerial Decisions*), as well as, 2 secondary ones (*Legislative Acts* and *Regulatory Provisions*). Nomothesia structures all legal documents, by using persistent URIs according to the template proposed by ELI http://www.legislation.di.uoa.gr/eli/{type}/{year}/{id}. For instance, for the first provision of the appointment example, a linking of Diavgeia with the Nomothesia URI http://legislation.di.uoa.gr/eli/law/2007/3549/article/25/paragraph/1 can be made. By integrating Nomothesia into Diavgeia, we also give citizens the ability to simply click to the legislative references of public sector decisions and see instantly the relevant passage of Greek legislation.

3.2 Metadata of Decisions

In addition to the uploading of the PDF file, public sector authorities also have to fill metadata information describing the decision. The metadata used vary according to the type of decision. For instance, the metadata of the *ExpenditureApproval* decision type holds important information about government's expenditure (such as the sender and receiver VAT registration numbers, the expense amount, etc). Diavgeia offers an OpenDataAPI (https://diavgeia.gov.gr/api/help) that can be used as an endpoint to query over the metadata. Despite OpenDataAPI being a step towards promoting transparency, inconsistency between decision text and

[8] https://sites.google.com/view/lddl-3/home

[9] The following is a fun example. A recent decision, listing the proposals that will be funded under a particular research and innovation call, starts with references to 33 (!) Greek laws. The good news was that 176 proposals were funded; one of them was our new project Choronomothesia which we discuss in Section 7.

metadata information is possible[10]. In our work, we embed metadata information into the RDF document, enforcing consistency.

3.3 Identifiers and Modifications of Decisions

Each decision is assigned a unique Internet Uploading Number (IUN), certifying that the decision has been uploaded on Diavgeia. IUN is important, since citizens and other public authorities can use decisions, by solely referring to their unique number. In addition to IUN, each decision is assigned a unique version token. Government institutions can upload a new version of a decision by claiming a new version token, but maintaining the same IUN. Diavgeia functions in an **append-only** manner, as it maintains the original decision with all its subsequent modifications, so it is amenable to blockchain technologies as we discuss in Section 5.

4 MODELING DECISIONS USING SEMANTIC WEB TECHNOLOGIES

In this section, we present an OWL ontology for modeling decisions of Diavgeia. We call our ontology *Diavgeia ontology* and we discuss its current version that adopts the ELI framework and the Nomothesia ontology. We present the Web editor and Visualizer components of DiavgeiaRedefined that generate and visualize the decisions expressed in RDF, respectively. Futhermore, we describe the linking of decisions with other datasets and we pose interesting SPARQL queries which take advantage of this interlinking.

4.1 The Diavgeia ontology

The ontology of Diavgeia is based on the pattern followed by public sector decisions, as discussed in Section 3.1. It imports and uses properties that are defined in the European Legislation Identifier (ELI) ontology and the Nomothesia ontology. The core[11] of Diavgeia ontology is shown on Figure 1. The 34 different decision types can be viewed as legal documents (class `LegalResource` of the ELI ontology). A decision (`LegalResource`) `changes` itself, by generating a new `version` and maintaining its unique `IUN`. A `LegalResource` consists of multiple `Considerations` and `Conclusions`. The `Consideration` class models the passages used to prove the validity of the decision (e.g., the three provisions of the appointment example), while `Conclusion` models the passages used as conclusions of the decision (that is the final passage of the appointment example). Both `Consideration` and `Conclusion` classes use the `cites` property of the ELI ontology to make a reference either to Greek Legislation (Nomothesia), or to another decision of Diavgeia. The `has_text` property describes the text body of either `Consideration` or `Conclusion`.

The Diavgeia ontology offers 121 properties to cover all the particularities of different decision types. In addition to `Consideration` and `Conclusion`, the ontology provides classes which describe important public sector activities. For instance, the class `Expenses` links an expense of a public authority to an individual or business. For the time being, this crucial information is expressed as metadata of the PDF decision, underlying the possibility of metadata inconsistency as described in Section 3.2. By merging metadata and decision text in a single RDF file, this possibility is eliminated.

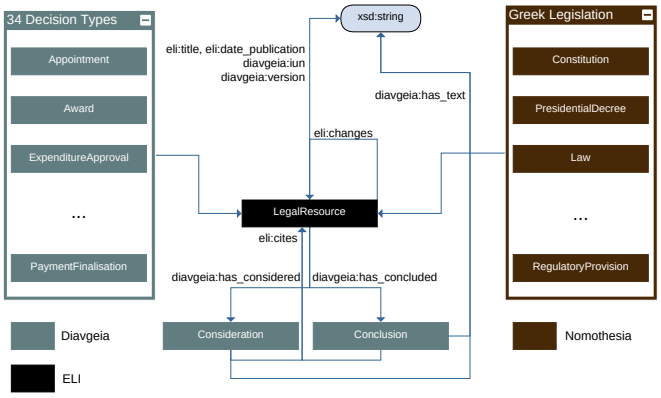


Figure 1: The core of Diavgeia ontology

In order to identify legal resources, we also need appropriate URIs. Persistent URIs is a strongly recommended best practice according to ELI. It is very important to have reliable means to identify the public sector decisions. Based on what is stated in Section 3.3, we can structure the persistent URIs of decisions according to the template http://www.diavgeia.gov.gr/eli/{iun}/{version}. Modifications of a decision result to the generation of a new URI which has the same `iun` and a new `version` number. Thus, the version of an enacted decision can be seen as the decision which has the most recent `date_publication` for a specific `iun`.

4.2 Web editor and Visualizer

DiavgeiaRedefined offers two main Web components in order to transparently adopt Semantic Web technologies to the production implementation of Diavgeia. The first one is a web editor for decisions, used exclusively by public sector authorities. The Web editor is a well-structured HTML form that government institutions can use to write their decisions. The HTML elements of the form are associated with the properties and classes of Diavgeia ontology. By submitting the form, the decision is stored as a compressed N3 file in the filesystem of Diavgeia and in Jena Apache's triple store.

The Visualizer is another component of DiavgeiaRedefined which can be used both by public authorities and citizens. Its purpose is to vizualize the decisions expressed in RDF inside a Web browser. Users provide the URI of the decision they want to visualize and the decision is displayed in the browser in a user-friendly manner.

4.3 Linking decisions with other public sector data

The linking of decisions with other public sector data, can be done by public authorities, using the Web editor component of DiavgeiaRedefined. Firstly, the Consideration or Conclusion classes of a decision, may make reference to the Greek Legislation of Nomothesia, as mentioned in Section 4.1. Linking Diavgeia with Nomothesia is easy, since the latter provides persisent URIs according to template http://www.legislation.di.uoa.gr/eli/{type}/{year}/{id}.

[10] An article about inconsistent metadata on Diavgeia website: https://eellak.ellak.gr/2016/07/06/veltionontas-tin-piotita-dedomenon-stin-diavgia/

[11] The full Diavgeia ontology is available on: https://github.com/ThemisB/diavgeiaRedefined/blob/master/rdf/diavgeia.owl

Public authorities have also to link `SpatialPlanningDecisions` decision type, with the dataset of administrative geography of Greece. Linking decisions with it is also easy and it is achieved through the construction of constant mappings.

4.4 Querying the Resulting RDF Data

By employing the Fuseki Server, we enable the formulation of complex queries over decisions of Diavgeia. This provides interested parties a mechanism to query the decisions of public sector organizations in a much easier way than in the current portal. The following query is such an example:

```
SELECT ?decision WHERE {
  ?decision diavgeia:has_expense ?expense;
            eli:date_publication ?date.
  ?expense diavgeia:expense_amount ?amount.
  FILTER (?date >= "2017-01-01"^^xsd:date &&
  ?date <= "2017-12-31"^^xsd:date)
} ORDER BY DESC(?amount) LIMIT 5
```

Retrieve the decisions taken in 2017 involving the 5 highest total amounts.

Our administrative colleagues from our university have told us that the same task in the current Diavgeia portal would take *months* to complete.

5 PRESERVING DIAVGEIA DECISIONS USING DISTRIBUTED LEDGERS

In this section, we describe the use of the bitcoin blockchain on Diavgeia. We present in detail the two blockchain tools that DiavgeiaRedefined offers, called Stamper and Consistency Verifier.

5.1 Stamper

Stamper is the tool which should be used by the administrators of Diavgeia in order to store public sector decisions on the bitcoin blockchain. The stamping procedure is described as follows:

(1) Government institutions upload their decisions on Diavgeia. The backend of Diavgeia stores decisions as compressed N3 files in its filesystem and in the triple store.
(2) The administrator of Diavgeia has to decide on the length of time intervals t for the stamping procedure, ensuring the integrity of decisions. Thus, the backend of Diavgeia starts a new stamping procedure every t time units.
(3) At the start of the stamping procedure, we find all the compressed N3 decisions which have not been stamped yet. Stamper organizes and aggregates these decisions into a Merkle Tree [7], using the hash function SHA-256. The root of the Merkle tree represents the fingerprint of the decisions which will be included in the forthcoming bitcoin transaction. By applying the SHA-256 hash function on the Merkle tree construction, the resulting root has a constant size of 32 bytes.
(4) The next step is to create a Bitcoin transaction and broadcast it to the rest of the network. DiavgeiaRedefined uses the bcoin library (http://bcoin.io/), offering Diavgeia an spv node[12], maintaining only a chain, a pool, and a *hierarchical deterministic (HD)* wallet [14] based on BIP44 [19].

A stamping transaction in our model consists of two outputs and one input. The first output contains the OP_RETURN opcode followed by the Merkle root in the *scriptPubKey* output (*scriptPubKey* = OP_RETURN + Root). This output guarantees the immutability of decisions. The second output is a pay-to-pubkey-hash[13], having as *pubKey* the next derived public address of the HD wallet. The input *scriptSig* consists of Diavgeia's signature and the current *publicKey* derived from HD wallet (*scriptSig* = *signature* + *publicKey*). The size of a stamping transaction is 267 bytes. In order to have certain guarantees that our transaction will be written into the next block and confirmed nearly immediately, mining fees can cost up to 120,150 satoshi (0.00125 bitcoin), which at the time of writing roughly amounts to $16.84.

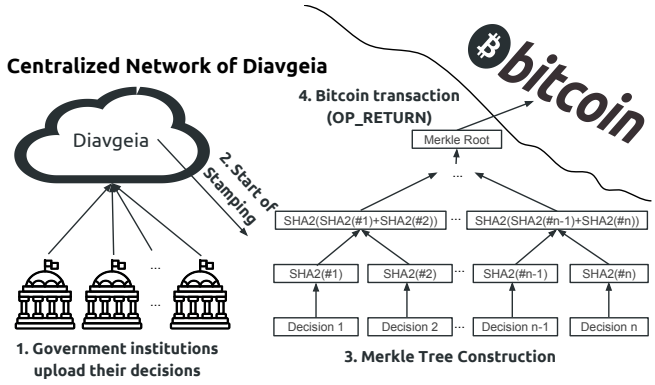


Figure 2: The Stamping procedure

After the end of each stamping transaction, Diavgeia publishes to its website the transaction identifier (Txid) and the order of decisions, as used for the Merkle tree construction. It also publishes once, the Master Public Key of its HD wallet. By publishing Diavgeia's master public key, interested parties are able to track the sequence of public keys and stamping transactions of Diavgeia. These publications are necessary to be made for the proper functionality of the Consistency Verifier (see Section 5.3).

5.2 Guarantees of Stamper

The Stamper tool provides high levels of immutability guarantees, especially when t value is configured to be small. Generally, the threat of a decision's modification or deletion appears on the time gap between two consecutive stampings. Small t values imply more stamping invocations and as a result Diavgeia creates more stamping transactions, but this comes at a higher cost. We consider a t value ranging from 3 hours to 1 day, to be an affordable solution for the government, since the daily cost of the usage of the blockchain will range from 0.00125 to 0.005 bitcoin ($16.84 - $134.72). The threshold for a decision's modification is also small, since an adversary (the administrators of Diavgeia, the government or other public authorities) are able to modify the decision in the next 3 hours to 1 day after its publishment.

[12] A method for verifying if particular transactions are included in a block without downloading the entire block (https://bitcoin.org/en/developer-guide#simplified-payment-verification-spv).

[13] https://en.bitcoin.it/wiki/Script#Standard_Transaction_to_Bitcoin_address_.28pay-to-pubkey-hash.29

As mentioned in Section 5.1, Stamper uses the open source bitcoin library (bcoin) to create the stamping transactions and relay them to the network. DiavgeiaRedefined does not use existing blockchain timestamping services (such as Stampery or OpenTimeStamps) because, in case of foul play, these services might be accused of having modified the Merkle root in the first place.

5.3 Consistency Verifier

Consistency Verifier is the tool used by the interested parties to verify the immutability of decisions over time. Algorithm 1 formalizes the steps it takes to verify the integrity of decisions.

Data: Decisions included in stamping transaction i: d_i, Master Public Key: mpk
Result: Boolean result of verification.
foreach $usedPublicAddress$ *of* mpk **do**
  $transaction \leftarrow$ getTransactionBySigScript($usedPublicAddress$);
  if $transaction$ *has OP_RETURN in the scriptPubKey output* **then**
    merkleTree $\leftarrow$ constructMerkleTree(d_i);
    if *merkleTree*$\rightarrow$*merkleRoot != *$transaction\rightarrow$*merkleRoot* **then**
      return false;
    end
  end
end
return true;

Algorithm 1: Consistency Verification procedure

The first step is to download the compressed N3 decisions which have been included in stamping transactions. Afterwards, the verifier downloads in ascending time order all bitcoin transactions (using the *chain.so* bitcoin block reader, available at https://chain.so/), related to the used public addresses derived from Diavgeia's master public key. In case of a stamping transaction, the verifier constructs the Merkle tree using the decisions of the first step. If the computed Merkle root is equal to the Merkle root found on the stamping transaction, decisions have remained unmodified.

6 EXPERIMENTAL EVALUATION

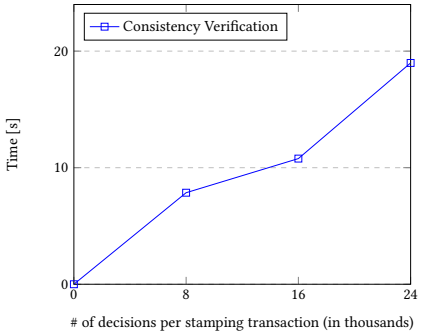
Here, we present a scalability evaluation of the Consistency Verifier tool, discussing the disk space reduction gzip compression of N3 files offers. Section 6.1 describes the synthetic dataset used in the Consistency Verifier experiment.

6.1 Dataset

To simulate the consistency verification process, we generated synthetic gzip N3 decisions[14], according to the *Diavgeia ontology*. Firstly, synthetic decisions have 7-17 `Consideration` and `Conclusion` class entities, each one of them has 150-350 random bytes as text part. Moreover, we have included several common-used properties, such as protocol number and thematic categories of a decision, as also information related to the departments of government institutions which upload them (phone number, address, etc).

We examine the time it takes an interested party to verify the consistency of Diavgeia in a month's common workload. We consider the scenario in which Diavgeia stores decisions on bitcoin blockchain, once a day. According to the Webpage of Diavgeia (https://diavgeia.gov.gr/en), the current rate of uploads is 16000 decisions per working day and assuming a month has 22 working days, we make 22 bitcoin stamping transactions. To examine the scalability of the verifier, we provide 3 different datasets, containing 8000, 16000 and 24000 decisions per day, summing up to 176000, 352000, 528000 compressed gzip N3 decisions, respectively.

[14]Datasets are available in: https://doi.org/10.6084/m9.figshare.5729292.v1



#of decisions per stamping tx	Verification Time (s)
8000	7.853
16000	10.780
24000	18.986

Figure 3: Evaluation of the Consistency Verifier

6.2 Test Environment

The verification experiment was run on a MacBook Pro with a 2.9 GHz Intel Core i5 processor and 8GB of memory, since this process may be executed by interested parties with a standard modern computer. The elapsed verification time is measured by Javascript's *console.time - console.timeEnd*. The execution time measures the time needed to create the 22 Merkle trees and compare the computed roots with the roots extracted from the stamping transactions. The recorded time does not take into account any network time; the time needed to download synthetic decisions from our Web server or to gather bitcoin transactions, by making requests to *chain.so*. To account for variability in the testing environment, the average of five independent runs is being reported.

6.3 Experimental Results

The experiment consisted of retrieving all synthetic decisions from our Web server and bitcoin transactions from *chain.so* and then compare the corresponding Merkle roots for validity, as presented in Algorithm 1. We use the 3 different datasets described in Section 6.1. The elapsed verification time is plotted in Figure 3.

This experiment validates the linear time growth of the Consistency Verifier. The integrity check for a month's regular workload, consisting of 16000 decisions per day, takes about 11 seconds. Even for the extreme case of 24000 decisions per day, the verifier takes approximately 19 seconds to perform the integrity check. These results validate the scalability of our blockchain solution and demonstrate that interested parties can efficiently perform integrity checks over the data of Diavgeia.

However, we acknowledge bitcoin's limitations in terms of cost, speed, and scalability [21]. Therefore, we would like to apply Stamper and Consistency Verifier to other blockchain technologies, like Ethereum [4].

6.4 Disk Space Reduction

Diavgeia currently hosts over 31 million decisions, leading to disk space limitations. The average size of a PDF-decision is about 2.5MB, summing up to a total of many TBs. We have created a sample, consisting of equivalent PDF and compressed gzip N3 files, for each different decision type of Diavgeia ontology[15]. For the sample, we find compressed N3 decisions about 86 times smaller, when compared to their PDF equivalent. Hence, decisions encoded in RDF allow for sophisticated SPARQL querying and also save space. In future work we also plan to use ideas from HDT [11] to lower storage requirements even further.

7 THE LINKED-DATA-ENABLED PORTAL NOMOTHESIA

Let us now present the Nomothesia platform which is the sister project of DiavgeiaRedefined.

The development of information systems archiving the content and metadata of legal documents is a common practice towards making legislation easily accessible to the public [5]. To name a few examples, the MetaLex document server [16] offers Dutch national regulations published by the official portal of the Dutch government, while the United Kingdom publishes legislation on its official portal. In the same spirit, [13] presents a service that offers Finnish legislation as linked data. Last, the Publications Office of the EU has developed a central content and metadata repository, called CELLAR, for storing official publications and bibliographic resources produced by the institutions of the EU [12].

Following in the footsteps of other successful efforts in Europe, we aim at modernizing the way Greek legislation is made public. We envision a new state of affairs in which ordinary citizens have advanced search capabilities at their fingertips on the content of legislation. We also envision that legislation is published in a way so that developers can consume it, and so that it can be also combined with other open data to increase its value for interested people. Currently, there is no other effort in Greece that takes this perspective on legislation and related decisions made by government institutions and administration alike.

Unfortunately, Greek legislation documents, published at the web site of the National Printing House of Greece[16], can be found mostly in PDF format, TXT formatted documents existing only for 2000 onwards. To make matters worse, while there are clearly defined instructions on how to write such a legal document, they are rarely followed, rendering any potential attempt to structure the legal text in a more manageable form particularly hard.

To address this, our group has developed a web platform called Nomothesia [6] (http://legislation.di.uoa.gr/), which consists of many components, each specifically dedicated to a proper task in order to create a pipeline of modern technologies. The idea is to automate the process of converting greek legal documents in a useful RDF knowledge graph any interested parties (and programmers) can use to extract useful information, or provide services with.

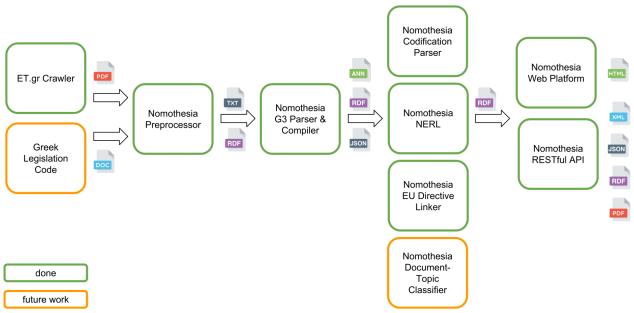


Figure 4: Nomothesia's Pipeline

The first phase of the pipeline, document processing, is achieved by crawling the web site of the National Printing House of Greece, gathering the documents, pre-processing and parsing them in order to produce structured RDF data; expanding upon ELI's ontology, each document is split into increasingly specific parts (books, parts, sections, chapters, articles, paragraphs, indents, lineas, citations), therefore allowing us to generate a complete structure of the original document. During this phase, metadata are also extracted from each document. Finally, two important aspects related to laws are addressed: its timeline and its codification.

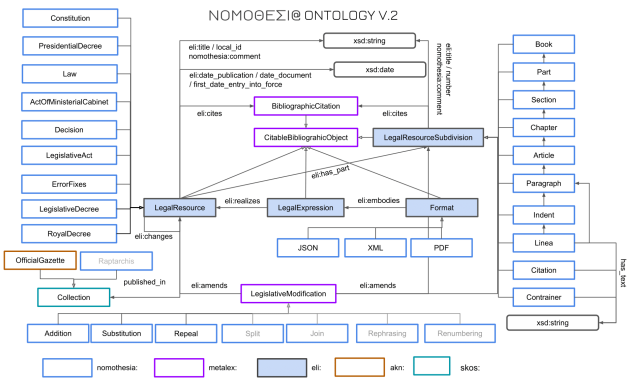


Figure 5: ELI's (expanded) ontology

There are 3 main dates involving a legal document: the date of publication, the date it was signed and the day it is set in force. While in many cases all 3 dates coincide with each other, there are some where they differ. When dealing with the temporal dimension of a law, it is very important to capture all 3 dates to do proper reasoning and inferencing. This is even more evident when tackling the codification of a law.

The codification basically represents all modifications a law has been subjected to, with major types being additions, substitutions and repeals. By having the 3 main dates of a law, it is possible to create a timeline where a law can be shown in the form it had during

[15]The sample is available on: https://github.com/ThemisB/diavgeiaRedefined/tree/master/rdf/samples

[16]http://www.et.gr/

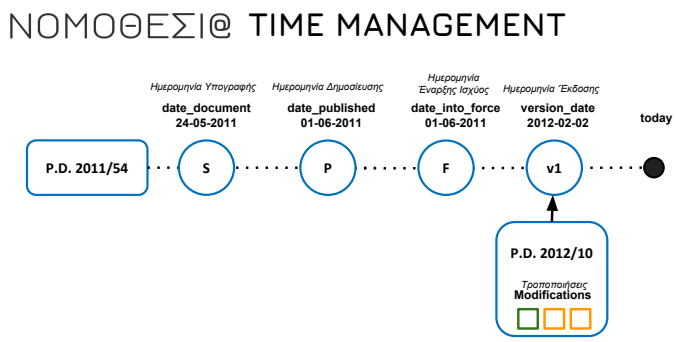


Figure 6: Nomothesia's timeline

a specific point in time. We take advantage of this knowledge in our web platform in order to showcase the legal text of a law in the exact form it is meant to have with all modifications applied to it up to the point in time the user wishes. Also, depending on the kind of modification in question, a different color is being used.

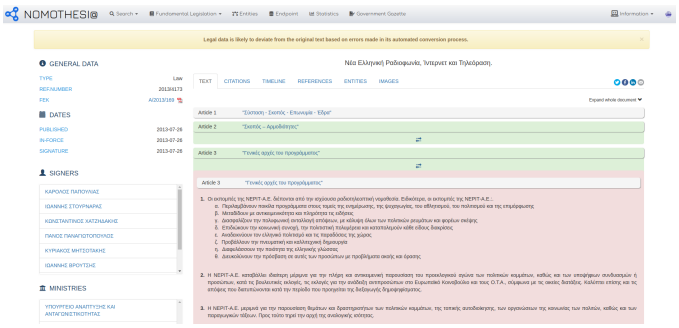


Figure 7: Nomothesia platform: visualizing law/2013/4173

The next phase addresses the information extraction and linking, both of which we thoroughly document in our work of [1]. Employing deep learning techniques, we predict for each token in some text its potential of being a reference to an entity. That reference is realized within a part of the text, but before it is of any use, it must be post-processed due to many abbreviations, cases and accents the greek language contains. When a normalized form is determined for each entity, we use the tool Silk [23] in order to interlink the current data with the third-party datasets of DBpedia Greek politicians, Greek administrative geography and European legislation. The final knowledge graph combines all that and, as a result, provides great querying capabilities.

Currently, Nomothesia contains about 5 million RDF triples which are interlinked in our knowledge graph. About 12,000 documents have been successfully codified, while our knowledge graph contains approximately 195,000 references to entities. We provide all legal documents of issues A and D of the National Printing House for the years 1990-2018, the penal and civil codes of Greece and all european directives and treaties extracted from EUR-Lex[17] in Greek. We provide persistent URIs for all rFesources of the knowledge graph, so developers can utilize RESTful API calls (we offer HTML, PDF, JSON, RDF, XML formats), as well as a SPARQL endpoint to allow direct querying on our knowledge graph, while we make the individual datasets we interlinked publicly available[18].

Currently, in collaboration with a company of landscape engineers in a Greek project called Choronomothesia, we aim to extract geospatial information as well, including maps, tables of coordinates etc. This is even harder than information extraction from text, because most maps or tables of coordinates are within images, occasionally handwritten and frequently maps are split among pages. This necessitates the usage of image analysis techniques to tackle the problem. However, extreme caution is required, because everything else is being extracted from text, while this information will be from images and, in order to end up with a meaningful knowledge graph, the two must be integrated. To the best of our knowledge, all the challenges faced by Choronomothesia are open problems in the relevant literature so far.

8 OTHER ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR INTELLIGENT PUBLIC ADMINISTRATIONS AND GOVERNMENTS

The previous sections discussed how linked data and distributed ledger technologies can help public administrations and governments become decentralized, trusted and linked with the aim of serving citizens in a more effective way. We now turn to discuss other Artificial Intelligence technologies that can also help towards the vision of making public administrations more *intelligent* than they are today. In our opinion, the following three intertwined technologies can offer very much in this direction:

- *Extracting information from legacy public sector documents.* Even if the Greek public sector decided to adopt DiavgeiaRedefined today, the decisions submitted as PDFs to the current portal in the last eight years are an untapped source of information. What can we do to extract useful information from these PDFs? We have already tackled this problem in Nomothesia platform using deep learning techniques [1] as we discussed in Section 7 already. Unfortunately, it is far from clear how to do this for Diavgeia. While it is challenging to address the intricacies of each issue of legal documents of the National Printing House in Nomothesia, as well as extracting information from PDF format, at least the documents' structure follows patterns that can make it easier to at least classify which documents would contain what information and in which potential places. In the documents of Diavgeia, this is not the case at all, as the documents are unstructured, each can contain classes of information others do not and, as a result, pose the challenge of finding a way to extract useful information from them reliably.
- *Knowledge graphs.* General knowledge graphs like DBpedia, Yago and Wikidata should be interlinked with public administration and government knowledge graphs. This will allow e.g., for sophisticated queries like "Which European countries invest in Computing and Artificial Intelligence less than what a United States technological university, based

[17] https://eur-lex.europa.eu/homepage.html

[18] http://legislation.di.uoa.gr/data

in Boston, is investing in a new college covering the same research area?". To answer this query, one has to combine information, e.g., from Wikidata, that shows that MIT will be investing 1 billion dollars to a new Computing and Artificial Intelligence college[19] and compare this amount with the total investment in the same research area by various European countries. The former information can probably be found in a general knowledge graph such as Wikidata, while the latter can be found in various government knowledge graphs.

- *Multilingual question answering*. Ordinary citizens and public sector employees do not speak SPARQL, although we wish they did; instead, they converse in natural language. Hence, current work in question answering over knowledge graphs and linked data [8] is fundamental in making public administration and government information more easily available to citizens. Work on question answering has evolved a lot in the last 10 years [8], but there are at least three kinds of questions that the currently available question answering systems do not deal with effectively: temporal questions (see TEQUILA for a nice start in this area [24]), questions involving quantities [24] and multilingual questions (see AMUSE [15] for a successful effort in this area). As an example, being able to deal with quantities is needed if one wants the question answering engine to be able to answer questions like the one mentioned in the previous paragraph, which involves finding information about a quantity (investment) and computing the total of such quantities.
- *Spoken dialogue systems.* When a citizen interacts with a public sector employee today, the interaction is through spoken dialogue. For example, if you request information about your submitted tax return in your local government office, the relevant public sector employee might ask for your tax identification number. Subsequently, if the answer to your enquiry was not the one expected, you might rephrase your question and ask again. Enter spoken dialogue systems! Spoken dialogue systems is a technology which can have immense applications in the public sector. Spoken dialogue systems, e.g., chatbots, can be there all the time, even late at night, when a citizen wants to request information about a transaction with the government e.g., information about one's tax return. To the best of our knowledge, spoken dialogue systems are not currently used in any part of the Greek public sector although their use, in the form of chatbots, is widespread in the private sector (e.g., in banks or telecom organizations etc.). However, chatbots are already in use in the public sector of other countries. For example, EMMA is a chatbot utilized by the U.S. Citizens and Immigration Services of the Dept. of Homeland Security.[20]
  Developing spoken dialogue systems is an important area in Artificial Intelligence research today, and advances to the state of the art in this topic are presented in many high quality conferences and journals, most prominently at SIGDIAL, the annual conference of the joint ACL/ISCA Special Interest Group on Discourse and Dialogue. The paper [20] which appeared in last year's SIGDIAL discusses spoken dialogue systems for navigating information expressed as linked data, and its ideas are certainly applicable to spoken dialogue systems for the public sector in Greece and elsewhere.

Our current efforts in both DiavgeiaRedefined and Nomothesia are concentrating on the above four research areas.

9 CONCLUSIONS

Let us now summarize what we learned from developing DiavgeiaRedefined and Nomothesia. Since we have already discussed technical improvements that we plan in these two systems in our previous papers [1, 2], this section focuses on the *social* aspects of our work, especially the lessons learned regarding the adoption of DiavgeiaRedefined and Nomothesia by the Greek public sector.

The first lesson that we learned is that, once you develop a system which disruptively effects in the public sector of your country and publish a couple of papers about it, you are not done. In fact, this is when the hard work of persuading the public sector to deploy an operational system based on your research and development starts. This is especially hard in a country like Greece where, in our opinion, the bureaucratic apparatus of the public sector typically fails to easily embrace innovative ideas coming from outsiders, including Greek universities. Our DiavgeiaRedefined and Nomothesia experiences are a case in point, as we explain above.

The first implementation of Nomothesia was completed by our group in September 2014. Since then, four more undergraduate theses and one M.Sc. thesis resulted in the creation of the current version of the portal. In April 2016, Nomothesia was presented and received an award in the 1st IT4GOV contest, which was organized by the Greek Ministry of Administrative Reform and Electronic Governance. Among the contest jury members were three ministers and people from the industry, who judged Nomothesia positively. Following this contest, we had been in contact with officials from the House of Parliament and the Ministry of Administrative Reform and Electronic Governance, who have expressed interest in deploying Nomothesia internally in the ministry. We have also presented Nomothesia to the Expert Group on European Legislation Identifier in Brussels on October 16, 2017. In all presentations, we have received very encouraging comments that motivated us to continue working on this topic. In March 2018, the Ministry of Administrative Reform (now without the "Electronic Governance" in its name) participated with us in a joint grant proposal in which DiavgeiaRedefined and Nomothesia would be the starting point of our joint work. The grant proposal was ranked very positively by the reviewers, but it was not selected for funding by the the funding agency due to the small budget of the call. In December 2018, we made a presentation to the Athens Bar Association about our platform. The association has an operational system called Isocrates [21] which is used by many lawyers in Greece today. In this meeting we explored various avenues for collaboration. In December 2018, we also made a detailed presentation to the Ministry of Administrative Reform (now under a new minister). We again stressed that we are happy to make the Nomothesia platform available to them for free, and that we think that both our research group and the ministry

[19] http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015

[20] https://www.uscis.gov/emma

[21] http://www.dsanet.gr/

would benefit from our joint collaboration. In general, through our interaction with the Greek public sector, we have realized that the Greek public sector is now ready to introduce Semantic Web and Linked Data technologies in the whole chain of law production which starts at the relevant ministry and ends with the approval of the law from the Hellenic Parliament, but there is no such readiness (or approved government policy) regarding the re-engineering of the transparency portal of Diavgeia.

The second lesson that we learned is that ordinary Greek citizens, with some knowledge of what Semantic Web, Linked Data and Distributed Ledger technologies can do for the public sector today, immediately understood the value of our proposals and were very supportive. Although we have received mostly positive feedback from people we did not personally know, we also received some constructive criticism. Some, for example, suggested that if the DiavgeiaRedefined ideas are adopted in the operational system of the Greek public sector, the only people to benefit will be Greek journalists working in the national yellow press who, together with a SPARQL expert, would then be able to pose queries to the portal and retrieve results much more easily. These results would then be interpreted in any way which fits the headlines that they want to have in their newspapers. This made us realize that the introduction of disruptive technologies, like Artificial Intelligence and Distributed Ledgers, in the public sector needs to be approached using a *participatory design approach* so that its impact becomes apparent to all stakeholders right from the start.

Finally, the third lesson that we learned is that it is important to teach Semantic Web and Linked Data technologies to graduate students in our universities. The desire to see Semantic Web and Linked Data technologies taught as part of the standard curriculum in Computer Science departments was also expressed by the participants of the recent panel "Enterprise Knowledge Graphs" which was organized by Natasha Noy at ISWC 2018. [22]

We close this paper with some advice to the Semantic Web and Linked Data community of which we are proud to be members. While the rest of the world seems determined to equate Artificial Intelligence with Deep Learning, we, as researchers working on Semantic Web and Linked Data, should stay alert and make a strong case for our technical contributions in the last 20 years that have resulted in robust technologies that are now in operation in many countries (especially in Europe) and in many applications. We should try to collaborate with researchers form other disciplines e.g., researchers from Distributed Ledger technologies and try to apply our technologies to areas where they can improve the lives of our fellow citizens. Our team will try to do this in the next three years in the context of the large European project AI4EU which involves 79 partners from academia, research and industry in Europe and which has a focus on *human-centered* Artificial Intelligence.[23]

10 ACKNOWLEDGEMENTS

The present work was co-funded by the European Union (ERDF) and Greek national funds through the Operational Program "Competitiveness, Entrepreneurship and Innovation", under the call "RESEARCH - CREATE - INNOVATE" (GeoREgC, project code: T1EDK 01000). We thank our colleagues Giorgos Apostolopoulos, Realdo Gatzi and Stavros Yannakis for their contributions to the most recent version of Nomothesia. Manolis Koubarakis would also like to thank Martin Kaltenböck, Pierre Maret and Stavros Vassos for interesting discussions regarding the Artificial Intelligence techniques discussed in Section 8 and their use in the public sector. Last, but not least, we would like to thank our colleagues Lydia Themeli, Olivia Katrakazi and Mirto Tsapatsari, who have spent an enormous number of hours fighting the Greek bureaucracy.

[22] http://iswc2018.semanticweb.org/panel-enterprise-scale-knowledge-graphs/
[23] http://ai4eu.org/

REFERENCES

[1] I. Angelidis, I. Chalkidis, and M. Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.
[2] T. Beris and M. Koubarakis. 2018. Modeling and Preserving Greek Government Decisions Using Semantic Web Technologies and Permissionless Blockchains. In *ESWC*.
[3] T. Berners-Lee. 2006. 5 star deployment scheme. https://www.w3.org/DesignIssues/LinkedData.html.
[4] Vitalik Buterin et al. 2014. A next-generation smart contract and decentralized application platform. *white paper* (2014).
[5] P. Casanovas, M. Palmirani, S. Peroni, T. M. van Engers, and F. Vitali. 2016. Semantic Web for the Legal Domain: The next step. *Semantic Web* 3 (2016).
[6] I. Chalkidis, Ch. Nikolaou, P. Soursos, and M. Koubarakis. 2017. Modeling and Querying Greek Legislation Using Semantic Web Technologies. In *ESWC*.
[7] J. Cucurull and J. Puiggali. 2016. Distributed Immutabilization of Secure Logs. In *STM*.
[8] D. Diefenbach, V. López, K. D. Singh, and P. Maret. 2017. Core techniques of question answering systems over knowledge bases: a survey. *Knowl. Inf. Syst.* (2017).
[9] M. English, S. Auer, and J. Domingue. 2016. Block Chain Technologies & The Semantic Web: A Framework for Symbiotic Development. In *CSCUBS*.
[10] J. G. Faísca and J. Q. Rogado. 2016. Decentralized Semantic Identity. In *SEMANTICS*.
[11] J. D. FernÃąndez, M. A. MartÃąnez-Prieto, C. GutiÃĺrrez, A. Polleres, and M. Arias. 2013. Binary RDF Representation for Publication and Exchange (HDT). *Journal of Web Semantics* (2013).
[12] E. Francesconi, M. Wilhelm Küster, P. Gratz, and S. Thelen. 2015. The Ontology-Based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. In *EGOVIS*.
[13] M. Frosterus, J. Tuominen, M. Wahlroos, and E. Hyvönen. [n. d.]. The Finnish Law as a Linked Data Service. In *ESWC 2013 Satellite Events*.
[14] G. Gutoski and D. Stebila. 2014. Hierarchical deterministic Bitcoin wallets that tolerate key leakage. *IACR* (2014).
[15] S. Hakimov, S. Jebbara, and P. Cimiano. 2017. AMUSE: Multilingual Semantic Parsing for Question Answering over Linked Data. In *ISWC*.
[16] R. Hoekstra. 2011. The MetaLex Document Server - Legal Documents as Versioned Linked Data. In *ISWC*.
[17] M. R. Hoffman. 2018. Can Blockchains and Linked Data Advance Taxation. In *WWW*.
[18] K. Janowicz, B. Regalia, P. Hitzler, G. Mai, S. Delbecque, M. Fröhlich, P. Martinent, and T. Lazarus. 2018. On the prospects of blockchain and distributed ledger technologies for open science and academic publishing. *Semantic Web* (2018).
[19] M. Palatinus and P. Rusnak. 2018. Multi-Account Hierarchy for Deterministic Wallets. https://github.com/bitcoin/bips/blob/master/bip-0044.mediawiki.
[20] A. Papangelis, P. Papadakos, Y. Stylianou, and Y. Tzitzikas. 2018. Spoken Dialogue for Information Navigation. In *SIGdial*.
[21] Manu Sporny. 2017. LD-DL'17 Workshop Keynote Talk By Many Sporny: Building Better Blockchains Via Linked Data. In *WWW (Companion Volume)*.
[22] A. Sutton and R. Samavi. 2017. Blockchain Enabled Privacy Audit Logs. In *ISWC*.
[23] J. Volz, Ch. Bizer, M. Gaedke, and G. Kobilarov. 2009. Silk - A Link Discovery Framework for the Web of Data *(CEUR)*.
[24] G. Weikum. 2018. Beyond Knowledge Graphs: New Frontiers of Machine Knowledge. The Web Conference Workshops.

Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud

Jeremy Debattista
debattij@tcd.ie
ADAPT Centre,
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

Judie Attard
attardj@tcd.ie
ADAPT Centre,
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

Rob Brennan
rob.brennan@dcu.ie
School of Computer Science
Dublin City University
Ireland

Declan O'Sullivan
declan.osullivan@scss.tcd.ie
ADAPT Centre,
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

ABSTRACT

The Linked Open Data (LOD) cloud has been around since 2007. Throughout the years, this prominent depiction served as the epitome for Linked Data and acted as a starting point for many. In this article we perform a number of experiments on the dataset metadata provided by the LOD cloud, in order to understand better whether the current visualised datasets are accessible and with an open license. Furthermore, we perform quality assessment of 17 metrics over accessible datasets that are part of the LOD cloud. These experiments were compared with previous experiments performed on older versions of the LOD cloud. The results showed that there was no improvement on previously identified problems. Based on our findings, we therefore propose a strategy and architecture for a potential collaborative and sustainable LOD cloud.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability; • **Information systems → World Wide Web**; *Web searching and information discovery*; *Semantic web description languages.*

KEYWORDS

Linked Data, LOD cloud, metadata quality, data quality, sustainable services

ACM Reference Format:
Jeremy Debattista, Judie Attard, Rob Brennan, and Declan O'Sullivan. 2019. Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3317075

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317075

1 INTRODUCTION

The Linked Open Data (LOD) cloud is regarded by many as a significant contribution towards the cause of adopting Linked Data(sets) and Semantic Web technologies, both in an academic context, but also to some extent in a commercial/industry one. Over time, the LOD cloud evolved into a *clustered catalog* of individual domain specific knowledge graphs. Despite this separation into various knowledge graphs, the LOD cloud demonstrates cohesion between these data sources with interlinks (links between datasets of a different domain), and intralinks (links between datasets of the same domain). The LOD cloud can be viewed by publishers as a catalog that can be crawled by data consumers for discovering datasets that can be re-used or linked to. For this to be possible, datasets on the LOD cloud has metadata attached to them, which its aim is to provide a level of understanding of how the data can be accessed and used.

The LOD cloud has been subject to a number of studies, especially on its metadata [2, 8, 15]. These studies highlight the shortcomings of the dataset metadata of the LOD cloud with regard to dataset accessibility, licensing, and metadata structure. These problems resulted in (1) dead links/data sources, (2) incorrect resolving of datasets, and (3) unclear usage of datasets in terms of licenses. These experiments were performed on previous versions of the LOD cloud, where the visualisation was less dynamic than lately; where monthly publication of the LOD cloud is currently performed. We therefore followed up on previous experiments to observe whether the problems with the metadata persist. Our results show that they do. We were therefore motivated to identify and discuss the current challenges going forward, and hence propose a sustainable architecture to help the Linked Data community to overcome these challenges. In a similar fashion, we also followed up on our previous quality assessment [8] survey in order to identify whether the

quality of accessible datasets on the LOD cloud has improved or otherwise.

The main contributions of this paper are therefore:

- An evaluation on the metadata of the LOD cloud (Section 2);
- A discussion of an on-going periodical data quality assessment over LOD cloud datasets on 17 different quality metrics (Section 3);
- Identification of a strategy and architecture for a potential collaborative and sustainable LOD cloud (Section 4).

We conclude our paper with related work in Section 5 and final remarks in Section 6.

2 THE DISCOVERABILITY AND OPENNESS OF DATASETS IN THE LOD CLOUD

Currently[1], the LOD cloud visualises 1,369 datasets[2]. The cloud diagram is updated every month with a tendency in increasing the number of dataset at each iteration. The inclusion criteria is based on the publishing of datasets following the Linked Data principles [3] and a set of five-rule inclusion guideline [13]. Furthermore, the "open" keyword suggests that datasets should also follow the open data definition, meaning that the dataset should "*be freely used, modified, and shared by anyone for any purpose*" [16]. Based on these premises, we conduct a number of experiments on the available metadata in order to understand better how these are enforced. Therefore, we will be looking at (1) what licenses are used, (2) if linked datasets are using the correct media types, and (3) the accessibility of the datasets. The size of the dataset and the number of external interlinks are disregarded in this experiment as these are related to the dataset itself. All experiments performed in this section are available online[3]. LOD cloud data can be retrieved from https://lod-cloud.net/lod-data.json and in this section we refer to this data as the *JSON data file* or the *data file*.

Experiments on dataset metadata have been conducted on previous versions of the LOD cloud [2, 8, 15]. In this paper we will compare the observed results to our previous work [8], where similar experiments were conducted. In [8], the observations were made in December 2015, based on the datasets that were linked and visualised in August 2014[4].

2.1 Can I freely use the dataset?

Licences are the heart of Open Data. They define whether third parties can re-use data or otherwise, and to what extent. In Linked Open Data, one would expect that such licenses are at least in a machine-readable format. Additionally, having the license mentioned in human-readable format, within the metadata's description, is a bonus. The open knowledge foundation recommend the usage of one of the following main licences to be in conformant with the open definition principles [16]:

- Creative Commons CCZero (CC0)
- Open Data Commons Public Domain Dedication and Licence (PDDL)
- Creative Commons Attribution 4.0 (CC-BY-4.0)
- Open Data Commons Attribution License (ODC-BY)
- Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)
- Open Data Commons Open Database License (ODbL)

For this experiment we parsed through each dataset in the data file and looked for the value attributed to `license` key. Our experiments show that around 45% (619 datasets) had a defined license. Figure 1 shows the frequency of the licenses used. The Creative Commons Attribution license is the most frequently used with 208 datasets or 15% using one of the recommended licenses. This is followed by the Creative Commons Attribution Share-Alike license (116 datasets) and Creative Commons Zero (89 datasets). 85 datasets are using the Creative Commons Non-Commerical license, which is not recommended by [16]. A more worrying aspect is that most of these licenses are links to human readable webpages which therefore cannot be understood by machine agents. In comparison to [8], our observations show an increase of 5% in the number of datasets that included a license in their metadata. We also noticed that whilst each dataset metadata can be exported to an RDF-based serialisation, the license is missing (e.g. https://lod-cloud.net/rdf/bio2rdf-taxonomy?format=ttl).

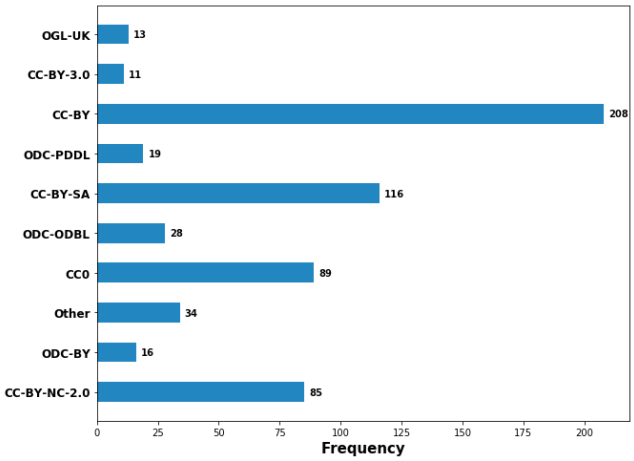


Figure 1: Overview of licenses used on the latest version of the LOD cloud.

We performed a second experiment in order to observe whether license data was included in the datasets' description field in a human-readable form. A regular expression that captures *license* or *copyright* and one of *under*, *grant* or *right* was performed on all 1,369 description fields. This resulted in 22 matches. We manually inspected the descriptions and observed that there were a total of 10 datasets that had a conformant license described, and 5 non-conformant licenses described. This gives an increase of 2 datasets from the observations in [8]. We also observed that there were 4 bad matches, meaning that these datasets were only listing the licenses of the data sources used to create the linked dataset. We also noticed that there were 3 datasets that had a particular license

[1]Based on the January 2019 crawl
[2]The website says 1,234 datasets, however, the official JSON file with the LOD cloud data indicates otherwise.
[3]https://github.com/jerdeb/lodexperiments
[4]For reference to the visualised LOD cloud diagram https://lod-cloud.net/versions/2014-08-30/lod-cloud.png. Last Accessed: 21st January 2019.

mentioned in the description but did not match the one used in the license field of the metadata.

2.2 Usage of the right media types for dataset distribution.

One of the main principles to add a dataset to the LOD cloud is that the datasets *"must resolve with or without content negotiation, to RDF data"* [13]. In this experiment we were interested in exploring the media types attached to the different distributions in the LOD cloud. Ideally, distributions in the LOD cloud use the respective media types, for example an RDF/XML data dump should use `application/rdf+xml`, as this would facilitate the uptake of the distributions by different agents by using the right content negotiation request. In Table 1 we list the different data types used for the distributions and their frequency. We observed a mixture of media types, however, the frequently used media type is `text/html`. Whilst this kind of media type is encouraged for RDFa serialisation, we observed that none of the distributions contained actual RDFa data. Similar to the findings in [8], we observed a number of unregistered media types. These included `RDF`, `n-quads`, `HTML`, and `application/x-ntriples` amongst others. 91 distributions had no media type assigned, whilst 109 distributions where assigned to media types that across the LOD cloud were less frequently used. We also noticed that there was a large number of distributions using `meta/void` (226 distributions) and `meta/rdf-schema` (370 distributions), however, these are also considered to be unknown media types.

Table 1: List of media types used in the distributions

Media Type	Frequency
mapping/owl	26
meta/owl	27
text/plain	31
application/x-gzip	32
n-quads	32
None	91
application/x-ntriples	91
meta/sitemap	102
application/x-nquads	103
Others	109
application/rdf+xml	114
application/octet-stream	118
HTML	119
application/zip	137
text/turtle	252
meta/void	266
meta/rdf-schema	370
RDF	401
text/html	1107

2.3 Accessibility of datasets

The final experiment on the datasets' metadata is to identify which datasets have a *potential* access point that allows for RDF crawling via an RDF dump, SPARQL endpoint, or a voID dataset description. We highlight the word *potential* because doing basic checks does not mean that the available distributions are well-formed RDF-based serialised datasets. For these three access points we set a number of criteria. The common criteria is that we first check if the access URL is online, with a 10 second timeout. For the data dumps the additional criteria is that the distributions have one of the following media types:

- `application/x-ntriples`
- `application/rdf+xml`
- `text/turtle`
- `application/x-nquads`
- `application/trig`
- `application/n-triples`
- `gzip:ntriples`
- `application/x-gzip`
- `application/octet-stream`
- `application/x-ntriples`
- `RDF`
- `plain/text`

For data dumps we added some invalid media types based on their frequency of use, however, this does not mean endorsement for wrong usage. With regard to SPARQL endpoints, the endpoint should answer to a simple `ASK` query, whilst the voID metadata should be able load in an in-memory graph structure and must answer return true to the query `ASK { ?s a <http://rdfs.org/ns/void#Dataset> . }`.

In contrast to [8] where it was observed that around 42% of the datasets had a potential direct access point from the LOD cloud, only 33% (454 datasets) of the datasets can potentially be accessed. Furthermore, from the 915 datasets that had no potential access points, 209 datasets had no data distributions, for example https://lod-cloud.net/dataset/slideshare2rdf, which violates the current LOD cloud inclusion criteria.

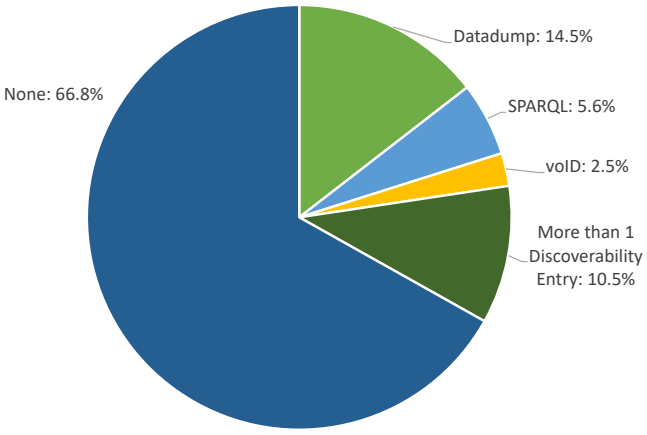


Figure 2: Visualisation of different access points for datasets.

Figure 2 illustrates the different access points. We can break this down as follows:

- Only Data dump - 199 datasets;
- Only SPARQL Endpoint - 77 datasets;

- Only voID - 34 datasets;
- Datadump and SPARQL Endpoint - 72 datasets;
- Datadump and voID - 62 datasets;
- SPARQL Endpoint and voID - 4 datasets;
- All three entry points - 6 datasets.

In Figure 3 we depict how the LOD cloud would look if only the datasets with a potential endpoint are visualised.

3 ASSESSING QUALITY ASPECTS OF REACHABLE DATASETS

Our next experiment focuses on the quality of datasets. Since August 2018, we performed monthly quality assessments on accessible datasets on the LOD cloud. The accessibility criteria used is the same as described in Section 2.3. In order not to re-download and pre-process data dumps[5], we monitor these for changes by querying the dumps URL header (`HTTP HEAD`) and check whether the `last-modified` or the `etag` values have changed. Unlike the accessibility metadata experiment, for quality assessment we also considered datasets that were only available via a LODLaundromat mirror. The aim of this experiment is to statistically observe datasets' quality, hence quality problems encountered are out of scope of this paper.

We used Luzzu [6], a Linked Data quality assessment framework, to assess the quality of the datasets on 17 different metrics from 3 different categories, mainly intrinsic, contextual, and representational. These metrics can be objectively assessed by any quality assessment framework, therefore limiting any bias that can arise from subjectively assessed metrics. The choice of metrics was left independent of any potential task, hence we understand that certain metrics might not be critical for different tasks, nonetheless, are important for the Linked Data community. On the other hand, the reason we have excluded accessibility category metrics from this experiment is due to the fact that we had a considerable number of datasets downloaded from LODLaundromat. Having a dataset crawled from LODLaundromat does not guarantee that the source is effectively online and hence would skew results. For a description and a formal definition of all metrics mentioned in this section, we refer the reader to [8]. The primary aim of this assessment was to create a service whereby data consumers can search for current LOD cloud datasets based on their quality. Furthermore, this assessment indicates whether the quality of linked datasets have improved or otherwise since the last assessment in [8]. The assessment in [8] was performed over 130 linked datasets. At time of writing we performed quality assessment on 451 datasets that could be directly accessed from the LOD cloud metadata (cf. Section 2.3). However, we also noticed that we had access to datasets that had very few or no triples. Therefore, prior to analysing the results we discarded the results of 71 quality assessments that were performed over datasets that had less than 100 triples. Overall, we assessed over 13B triples, and all downloaded and assessed datasets are available online as HDT dumps[6] and GZ dumps[7].

Figure 4 shows how the average quality changed between the previous assessment and over the 6 month period August 2018 and January 2019 for the representational category (6 metrics) and contextual category (2 metrics) metrics. The chart also demonstrates the number of datasets assessed at each observation date. The metrics assessed in these two categories are:

- (RC1) Keeping URIs Short
- (RC2) Minimal Usage of RDF Data Structures
- (IN3) Usage of Undefined Classes and Properties
- (IN4) Usage of Blank Nodes
- (V1) Different Serialisation Formats
- (V2) Usage of Multiple Languages
- (P1) Provision of Basic Provenance Information
- (U1) Human Readable Labelling and Comments

In most cases we observe that the average has increased between the 2016 assessment and the latest 6 month assessment. Given that the number of datasets has increased drastically, we also calculated the standard deviation for the different months and compared them with [8]. We observed that the spread of the values of all metrics were similar (± 5% standard deviation points), with the exception of IN4. The observation of August 2018 for IN4 had a standard deviation of 2.89% (median value of 99.95), when compared to 12.15% of the 2016 assessment. However, the standard deviation increased to around 6.66% for the following months. We observe that the datasets obtained in the period August 2018 and January 2019 have improved on metrics RC1, IN3, P1, and U1. With regard to P1 and U1, we observe an average increase of 7.7% and 7.2% points respectively, whilst for RC1 we saw an average of 5.9% increase, whilst for IN3 an average of 8.5% increase. There was no significant change with regard to the V1 and V2 metrics, which are not displayed in the chart as their value is an integer value in contrast to the rest which are percentages.

Figure 5 shows how the average quality values changed from the 2016 assessment and over the 6 month period for the intrinsic category metrics. The metrics assessed in this categories are:

- (CN2) Extensional Conciseness
- (CS1) Entities as Members of Disjoint Classes
- (CS2) Misplaced Classes or Properties
- (CS3) Misused OWL Datatype or Object Properties
- (CS4) Usage of Deprecated Classes or Properties
- (CS5) Valid Usage of the Inverse Functional Property
- (CS6) Ontology Hijacking
- (CS9) Usage of Incorrect Domain or Range Datatypes
- (SV3) Compatible Datatype

For the intrinsic metrics the situation seems more balanced. We observed an increase in metric CS9 or around 9.16% when comparing the first assessment against the 6 consecutive months, however the spread of the data points is similar. On the other hand, we saw a decrease in quality for the CN2 and CS6 metrics. The average decrease is of 4.24% and 6.47% respectively. Furthermore, we noticed that the standard deviation for CS6 increased by around 10% (median 100%), thus the average value would ultimately be affected by the less conformant datasets. Metric SV3 also saw a slight increase on average, however, we also noticed that the spread of quality among datasets in this regard is much less (on average 3.7% vs 14.6%, with both medians 100%) than in [8]. Therefore, from a statistical point of view datasets assessed between August 2018 and January 2019 tend to be more conformant towards this aspect than CS6.

[5] We follow the pre-processing step described in [8, §4.1]
[6] http://s001.adaptcentre.ie/lod/hdtdumps/
[7] http://s001.adaptcentre.ie/lod/gzdumps/

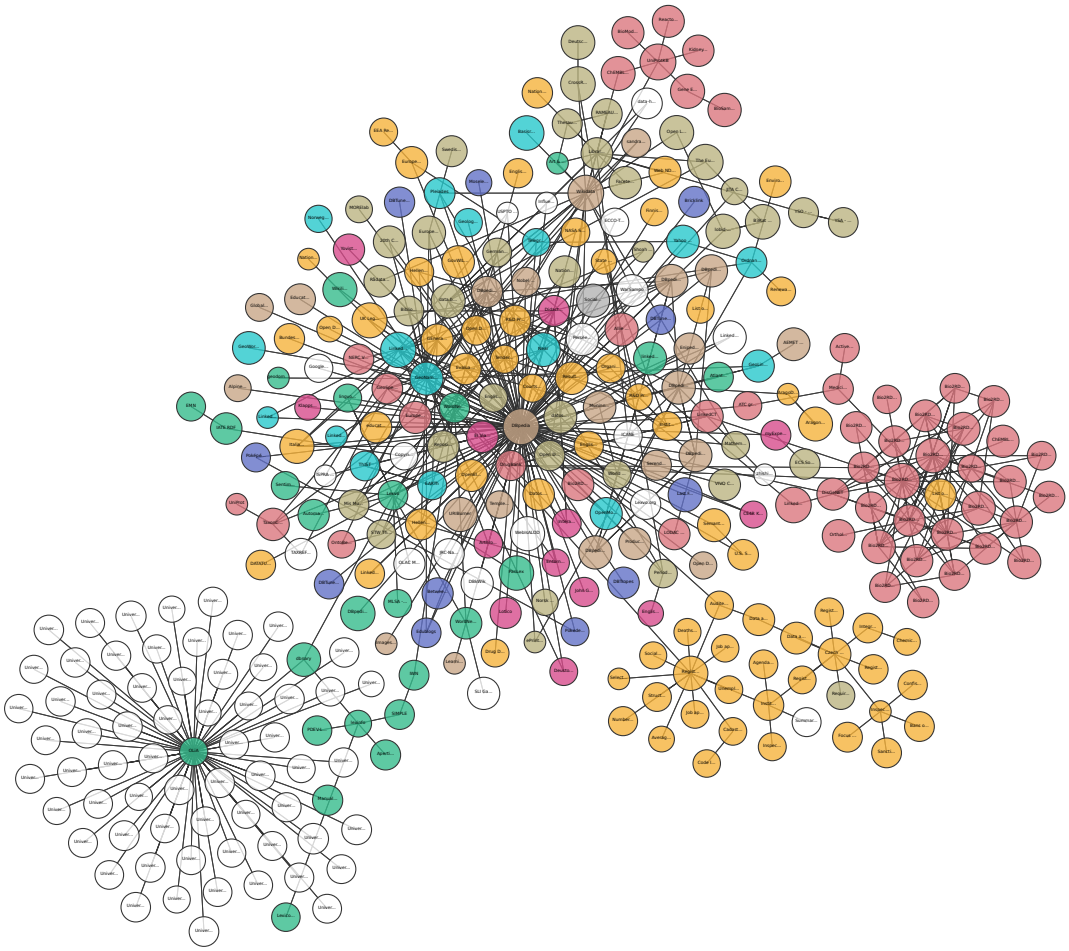


Figure 3: Datasets in LOD with a potential access point. The colours refer to domains as identified by the LOD cloud maintainers.

All quality metadata results and the crawled datasets used are available in an prototype online catalog: http://luzzu.adaptcentre.ie. The metadata schema is based on the Dataset Quality Vocabulary [7] (daQ), which is part of Luzzu's underlying semantic framework [6]. Unfortunately, we do not yet have enough data to potentially predict how the quality could look like in a number of years, however, we plan to keep our monthly crawls to gather more observations.

4 THE SUSTAINABLE LOD CLOUD

The current LOD cloud poses a number of barriers when data consumers are trying to identify a particular RDF dataset. Previous work [2, 8, 15] has already highlighted problems related to generated dataset metadata, however, as we discuss in Section 2, these have not been solved. Given the increase of RDF-based linked datasets on the web, the Linked Data community should ensure that the LOD cloud does not end up dormant and outdated, as happened between 2014 and 2017. Therefore, in order to make the LOD cloud a sustainable service and to provide a motivation for a new approach, we have to tackle the following challenges:

C1 - Publishers should own and maintain the datasets' metadata.
The LOD cloud is perceived as a monolithic structure with dataset metadata stored in a centralised catalogue. The metadata is initially created by the publishers themselves, however, changing this metadata is more difficult since this data is now "owned" by the LOD cloud maintainers. Therefore, whilst datasets might be updated, the metadata in the LOD cloud might still be old.

C2 - Lack of systematic and fine-granular metadata structures.
One of the major problems or challenges in the current state of the LOD cloud is that the metadata has no systematic structure in terms of properties, the property's values, and categorical values (e.g. media type). The current LOD cloud metadata attempts to leverage on both DCAT [12] and voID [1] standards, using predicates from both vocabularies to achieve a granular metadata description. This challenge

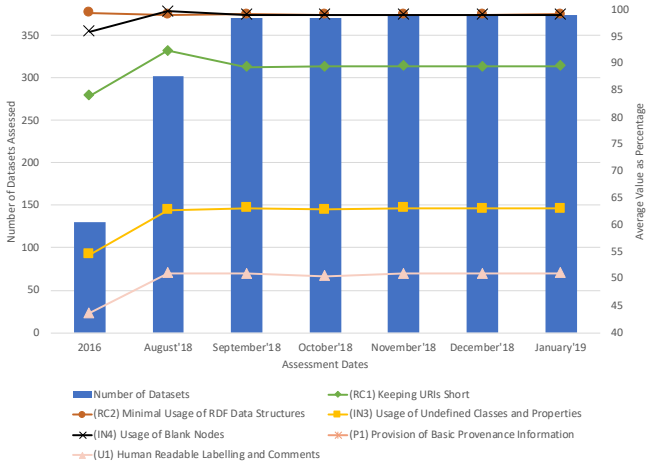


Figure 4: Aggregating quality metric results for representational and contextual metrics.

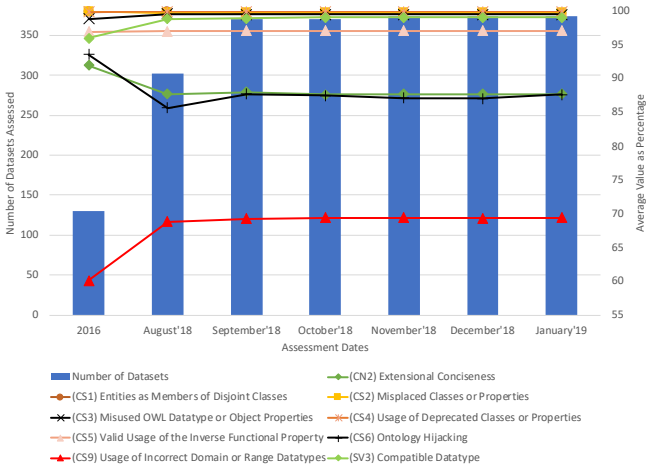


Figure 5: Aggregating quality metric results for intrinsic metrics.

is highly coupled with the next challenge, C3, which looks at the validation of metadata descriptions.

C3 - Invalid metadata descriptions.
Giving ownership of the metadata back to the publishers is no a guarantee that the metadata will be updated or factual. Therefore, validators should be implemented. The current LOD cloud suffers from metadata validation. Manually inspecting the DBpedia metadata [8] as an example, we found that the `void:dataDump` predicate links to the DBpedia downloads page. This is incorrect as the editors of the voID vocabulary documentation highlights that "*the* `void:dataDump` *property should not be used for linking to a download web page ... [it] should only be used for linking directly to dump files*" [1]. Furthermore, the license points to the a web page that describes the license, rather than the actual license itself (http://opendefinition.org/licenses/cc-by-sa/), whilst among the different media types attached to the dataset's distribution, we found `HTML`, `meta/void`, `linked data`, and `RDF`.

C4 - Many dead and outdated datesets listed.
Datasets that have perished from the web are still being visualised on the LOD cloud. The eleven 270a.info [4] datasets[9] are a notable example. In this case, we observe that an archived partial snapshot has been stored in LOD Laundromat[10], however, the metadata still shows the original description. Whilst archiving these datasets for preservation using tools such as LOD Laundromat is important, these datasets should not be visualised in the LOD cloud.

C5 - Lack of involvement of data consumers in the structure.
Up till now, data consumers could, painfully, crawl or search the LOD cloud diagram for a potential dataset. Prior to the 2018 update, most crawling or searching was done either by using datahub.io API or else by parsing the visualised SVG diagram. The latest versions provide a JSON file with all the datasets. Given the assumption that RDF/Linked Data consumers are potentially the largest set of users using the LOD cloud, the current versions still lack adequate filtering and searching tools, and lessons could be learned from the recent Google Dataset search[11] portal.

These five challenges guide us throughout this section to propose (1) a set of sustainability strategies for the LOD cloud, and (2) a potentially sustainable architecture based on Linked Data principles.

4.1 Sustainability strategies for the LOD cloud

The first step that is required in order to make the LOD cloud sustainable is to define a strategy. This strategy is based upon *people* or stakeholders, *processes* and *technology*. In a nutshell, the stakeholders of the LOD cloud service are the dataset publishers, the service maintainer/s, third party service providers, and the data consumers. The main process is the publishing of datasets on the LOD cloud using standards and an interoperable data model. Finally, the technology ensures the automation of adding and validating the status of datasets on the LOD cloud.

Based on Collibra's[12] experience with regard to building data governance solutions [14], we adapted their best practices as guidelines for the proposed LOD cloud service strategy as follows:

(1) **The service operating model.**
Challenge(s) to be tackled: C1, C5
Whilst acknowledging the previous and current maintainers, we propose that the LOD cloud service is operated on a federated model ensuring that the cloud remains open and the responsibility of the community as a whole. However, this requires instilling a culture that encourages interaction between the various stakeholders. *Publishers* should provide dataset metadata and be responsible to maintain it, ensure

[8] https://lod-cloud.net/dataset/dbpedia

[9] One such example: https://lod-cloud.net/dataset/fao-linked-data
[10] http://download.lodlaundromat.org/2a3bed796c47b67919645 9f3b5612b65
[11] https://toolbox.google.com/datasetsearch
[12] https://www.collibra.com

that their datasets are of high quality, and guarantee high availability uptime of their data services (e.g. SPARQL endpoints). The *LOD cloud service maintainer/s* must control services related to the generation, cataloging (i.e. keeping a list of all submitted and verified dataset metadata resource URIs), availability and maintenance of the cloud. *Consumers* should be able to comment (e.g. on usage) and vote on different data sources, helping potential future consumers to decide on whether a data source is right for their use case. Finally, *third party service providers*, such as quality assessors, should be able to identify changes in the LOD cloud and generate output that can be easily linked to the datasets' metadata in the LOD cloud.

(2) **Identification of critical data elements.**
Challenge(s) to be tackled: C2
In order to have a uniform view of the datasets in the LOD cloud, the critical element is the identification of a metadata standard, and a glossary or taxonomy for non-descriptive values (e.g. for licenses, or media types). Over the years, standards and taxonomies have been defined and based on Linked Data principles we advocate for reuse where possible. Datasets' metadata should clearly identify the ownership and usage of the dataset, providing: (a) *who* owns the data; allowing consumers to identify whether the publisher is a trusted source that can be reached to answer any questions; (b) *what* the dataset is about; including a basic description, purpose of the dataset, and the schemas underlying the data; (c) *where* could dataset distributions and/or other data services such as endpoints be found; and (d) *how* can the dataset be used in terms of licensing. Whilst most of this is already catered for in the current LOD cloud service, and improved over the years (for example using drop down lists in forms with predefined values) there is no agreed upon glossary amongst the publishers and the service maintainers themselves.

(3) **Defining the key activities and control structures for sustainability of service.**
Challenge(s) to be tackled: C3, C4
In order to define control measurements for the LOD cloud service, we first need to identify the key workflow activities. There are a number of key activities between the different stakeholders mentioned previously that would require different workflows. The first key workflow in the LOD cloud service is to check the validation and the correctness of a submitted dataset metadata resource URI by a publisher. The metadata resource URI is dereferenced as required by the Linked Data principles, and validated prior to inclusion in the LOD cloud. Another key workflow between the LOD cloud service and the publisher is the provision of regular heartbeat checks to ensure the availability of subscribed datasets for data consumers. Allowing for data consumers to vote and comment on specific data sources requires an authority or filtering from the LOD cloud service. A third workflow is a mechanism that prevents abuse or spam, ensuring sanity checks on subjective views of the datasets. This ensures that potential future consumers are not misled in using or disregarding a particular dataset or publisher. Similarly, the LOD cloud service should guarantee that external service providers are not biasing towards particular datasets or publishers, and any output that is generated (e.g. quality metadata), should contain extensive data lineage and provenance information, providing traceability to all stakeholders. Trustworthiness between the stakeholders, of the dataset, and other metadata generated by the different service providers, and the sustainability of the service, depends on the successful implementation of these control structures.

4.2 Capabilities and architecture of a LOD cloud service.

Based on the identified strategies, in this section we define the capabilities of the proposed LOD cloud service. The architecture captures the three pillars of the strategy, that is the *people*, the *processes* and the *technology*. The architecture is driven on the following capabilities:

Discovery - The LOD cloud service must enable data consumers, in this case both people and agents, to search, explore and identify the datasets required for a particular task in the most quick and efficient manner.

Understandability - The LOD cloud service is built on top of an interoperable semantic layer of standards and taxonomies that all publishers use uniformly. Furthermore, the external service providers must provide any output data using the same data model as in the LOD cloud service, i.e. RDF.

Social - The LOD cloud service allows all stakeholders to participate in the upkeep and uptake of the LOD cloud.

4.2.1 Architecture. Based on the idea of the LOD Research cloud[13], the underlying LOD cloud service should run using a Linked Data platform, which accepts Linked Data Notifications [5]. The idea is that once publishers publish their metadata resource on their servers, the publisher sends a Linked Data notification with the resource. Listing 1 illustrates a sample publishers' notification that is sent to the LOD cloud service. Upon receiving the notification, the LOD cloud service dereferences the received request and validates the dataset's metadata (Section 4.2.2). Once validated, the metadata's URI is stored in a registry and eventually visualised as part of the LOD cloud. Furthermore, the LOD cloud service will perform regular heartbeat checks on the publishers. Service providers and data consumers can then directly communicate with the publisher by dereferencing the metadata's URI that is stored in the LOD cloud service. Any output from the service providers and consumers (e.g. votes or comments) can be stored on the LOD cloud service triple store. Examples of service providers might include frameworks for quality, value, profiling, archiving amongst others, producing metadata that can be linked to datasets' metadata resources subscribed within the LOD cloud. Human consumers should be able to filter and search datasets based on various criteria, allowing them to vote or comment on these datasets, hence filtering mechanisms should be in place. Figure 6 depicts a high-level architecture diagram of the proposed solution.

```
@prefix schema: <http://schema.org/> .
[   a schema:CreateAction;
     schema:agent [
```

[13] https://linkedresearch.org/cloud. Last Accessed: 1st February 2019.

```
      a schema:Organisation;
      schema:name "The Org"
    ];
    schema:object <http://the.org/metadata/thedataset>
] .
```

Listing 1: Publisher's Linked Data Notification.

4.2.2 Dataset Metadata. To date, adding a dataset to the LOD cloud was either by adding dataset metadata directly to datahub.io with the LOD cloud tag, or more recently by filling out a form with all required fields. The former resulted into DCAT metadata, whilst for the latter form data is mapped into a combination of voID and DCAT metadata. Whilst ideally all publishers will publish their dataset metadata using one agreed upon standard, the LOD cloud ecosystem should allow any metadata published either by DCAT, voID or schema.org. The mandatory fields in the metadata are:

- **Title** - The dataset's title in textual form;
- **Description** - A concise textual description of the dataset which might also include information such as how it was used and a human-readable license;
- **Creator** - A resource which describes the creator or the publisher of the dataset, whom consumers should contact for questions related to the dataset;
- **Website** - A human-readable web page describing in more detail the dataset;
- **Full Download** - A resource which describes the distribution of the full data dump of the dataset. Each distribution must have the media type (mime type) described, for example for and RDF/XML data dump, one should use "application/rdf+xml". Furthermore, one should use the Media Types as Linked Data resources [17], which contains semantic descriptions for different RDF-based serialisation This is not mandatory if a SPARQL endpoint is provided;
- **SPARQL Endpoint** - Similar to full download, a resource that describes the access for the SPARQL endpoint, which might or might not include the different SPARQL protocols available. This is not mandatory if a full download is provided;
- **Domain** - A textual description of the domain of the dataset, for example financial or geospatial; and
- **License** - A machine-readable resource that describes the legality of reuse of the given dataset (dump or endpoint). Usage of correct machine-readable licenses such as [18] or creative commons semantic URIs are mandatory.

Other fields such as DOI, example resources, data catalogue, number of triples, and links could also be provided. Nonetheless, the LOD cloud services should identify the linked and number of triples for each subscribed dataset automatically.

This paper contributes towards an evaluation of the LOD cloud metadata, a periodical quality analysis of linked datasets that can be crawled from the LOD cloud, and propose strategies and an approach for making the LOD cloud service more sustainable.

5 RELATED WORK

Schmachtenberg et al. [19] crawled the Web of Data in order to present the 2014 version of the LOD cloud diagram. The authors analysed how different best practices were adopted by the crawled datasets, more specifically looking at provenance, licensing, and access methods amongst other experiments. In a similar crawling approach, Neto et al. [15] attempted to identify the actual Linked Open Data cloud by obtaining metadata from different sources. One of their goals was to assess the quality of the available RDF metadata. Similar to our findings, the authors report that the metadata suffered from a number of quality issues, mainly, lack of usage of standard vocabularies, incorrect usage of specific properties, and erroneous data. Assaf et al. [2] discussed the quality of the metadata of datasets available in the 2014 version of the LOD cloud. For this, the authors made use of the datahub.io API, thus using datahub's provided metadata as their evidence. Assaf et al. concluded that the metadata is in bad condition, mostly as a result of noisy data in licensing and accessibility metadata fields.

More recently, Debattista et al. [8] performed a study on the quality of the metadata and datasets available on the 2014 version of the LOD cloud. With regard to metadata quality, the authors conclude that based on the Open Defintion [16], approximately less than half of the datasets should not be part of the LOD cloud. The authors also suggested that the LOD cloud should reflect the Web of Data and be more dynamic. In terms of data quality, Debattista et al. assessed 27 metrics over 130 datasets. The authors concluded that the resulting overall aggregated average of slightly below 60% indicated that linked datasets might have a better quality than perceived [9]. In earlier studies, Hogan et al. [10] crawled and assessed the quality of around 12 million RDF statements. The aim of this study was to find and discuss common problems related to accessibility, reasoning, syntactical and non-authoritative contributions. In a follow up study [11], Hogan et al. assessed the quality of over 1 billion quads on a number of best practices and guidelines.

While a number of literature analysed the LOD cloud and metadata of linked datasets, to the best of our knowledge there is no work discussing how to concretely mitigate the identified problems. This motivate further our aim to propose a sustainable strategy for the LOD cloud service.

6 CONCLUSIONS

The first part of the article's title asks the question *Is the LOD cloud at risk of becoming a museum for datasets?* We have performed a number of experiments on the LOD cloud metadata on its structure, access and licensing parts. Values from these gave us a clear indication on the current status. We observed that there are datasets that have been offline but are still visualised, datasets that have incorrect or no license, and moreover datasets that do not adhere to the 5 LOD cloud inclusion principles. Furthermore, being Linked *Open* Data, one would expect that datasets follow the open definition [16] or the well known Linked Open Data 5 star. When we create the intersection of the dataset that have an access point and an open licence, we ended up with 35 datasets that can be called linked open datasets. On the other hand, the overall quality of datasets generally improved for the majority of datasets. Nonetheless, data publishers should invest in having quality checks within their publishing frameworks. Finally, when compared with similar experiments performed on earlier version of the LOD cloud, we can conclude that this significant contribution is in a great risk of becoming a museum for *linked* (partially *open*) *datasets*.

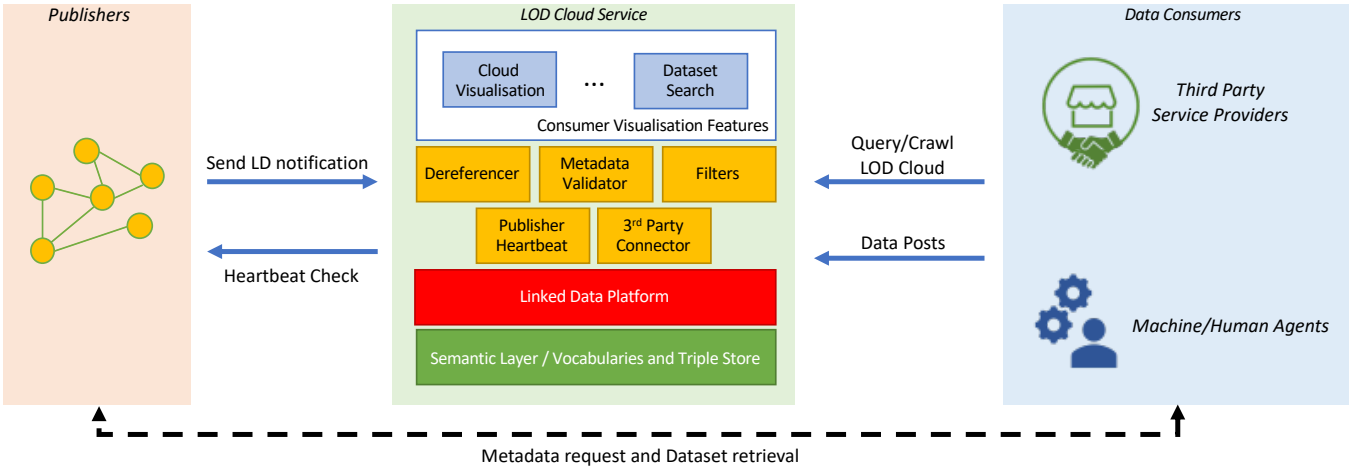


Figure 6: Proposed High-Level Architecture Diagram of the LOD cloud Service and Principle Data Flow.

Striving for the survival of the LOD cloud, in this article we propose a sustainable strategy for a new LOD cloud service, based on all stakeholders of the LOD cloud and standard Linked Data vocabularies and technologies. We discussed challenges that were identified in both previous literature and this study. Finally, we described the capabilities and an potential architecture for the LOD cloud service. As for future work, we encourage more discussion with the various stakeholders on how to save the LOD cloud from its untimely death, create a first prototype of the proposed architecture, and finally foster a culture of collaboration between all stakeholders of the LOD cloud.

ACKNOWLEDGMENTS

This research was partially supported by the Irish Research Council Government of Ireland Postdoctoral Fellowship [project ID GOIPD/2017/1204], Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (EDGE) [grant agreement no. 713567], and Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106].

REFERENCES

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 2011. *Describing Linked Datasets with the VoID Vocabulary.* W3C Interest Group Note. World Wide Web Consortium.

[2] Ahmad Assaf, Raphaël Troncy, and Aline Senart. 2015. What's up LOD Cloud? - Observing the State of Linked Open Data Cloud Metadata. In *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers (Lecture Notes in Computer Science)*, Fabien Gandon, Christophe Guéret, Serena Villata, John G. Breslin, Catherine Faron-Zucker, and Antoine Zimmermann (Eds.), Vol. 9341. Springer, 247–254. https://doi.org/10.1007/978-3-319-25639-9_40

[3] Tim Berners-Lee. 2006. Linked Data – Design Issues. https://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2017-12-15.

[4] Sarven Capadisli, Sören Auer, and Axel-Cyrille Ngonga Ngomo. 2015. Linked SDMX Data: Path to high fidelity Statistical Linked Data. *Semantic Web* 6, 2 (2015). https://doi.org/10.3233/SW-130123

[5] Sarven Capadisli and Amy Guy. 2017. *Linked Data Notifications.* W3C Recommendation. World Wide Web Consortium (W3C).

[6] J. Debattista, S. Auer, and C. Lange. 2016. Luzzu – A Methodology and Framework for Linked Data Quality Assessment. *Data and Information Quality* 8, 1 (Oct. 2016).

[7] Jeremy Debattista, Christoph Lange, and Sören Auer. 2014. Representing dataset quality metadata using multi-dimensional views. In *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, Harald Sack, Agata Filipowska, Jens Lehmann, and Sebastian Hellmann (Eds.). ACM, 92–99. https://doi.org/10.1145/2660517.2660525

[8] Jeremy Debattista, Christoph Lange, Sören Auer, and Dominic Cortis. 2018. Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web* 9, 6 (Sep 2018), 859–901. https://doi.org/10.3233/SW-180306

[9] Pascal Hitzler and Krzysztof Janowicz. 2013. Linked Data, Big Data, and the 4th Paradigm. *Semantic Web* 4, 3 (2013), 233–235.

[10] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. 2010. Weaving the Pedantic Web. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010 (CEUR Workshop Proceedings)*, Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas (Eds.), Vol. 628. CEUR-WS.org. http://ceur-ws.org/Vol-628/ldow2010_paper04.pdf

[11] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. 2012. An empirical survey of Linked Data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web* 14 (jul 2012), 14–44. https://doi.org/10.1016/j.websem.2012.02.001

[12] Fadi Maali, John Erickson, and Phil Archer. 2014. *Data Catalog Vocabulary (DCAT).* W3C Recommendation. World Wide Web Consortium. http://www.w3.org/TR/vocab-dcat/

[13] John P. McCrae, Andrejs Abele, Paul Buitelaar, Richard Cyganiak, Anja Jentzsch, and Vladimir Andryushechkin. 2019. Linked Open Data Cloud. http://lod-cloud.net

[14] Kash Mehdi. 2017. 4 Data Governance Best Practices to Kickstart your Data Governance Program. Blog Post. https://www.collibra.com/blog/4-data-governance-best-practices Last Accessed: 2019-01-24.

[15] Ciro Baron Neto, Dimitris Kontokostas, Amit Kirschenbaum, Gustavo Correa Publio, Diego Esteves, and Sebastian Hellmann. 2017. IDOL: Comprehensive & Complete LOD Insights. In *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, Rinke Hoekstra, Catherine Faron-Zucker, Tassilo Pellegrini, and Victor de Boer (Eds.). ACM, 49–56. https://doi.org/10.1145/3132218.3132238

[16] Open Knowledge Foundation. [n. d.]. The Open Definition. http://opendefinition.org. Accessed: 2017-12-15.

[17] Silvio Peroni. 2016. Media type as Linked Open Data. http://www.sparontologies.net/mediatype/.

[18] Victor Rodriguez-Doncel, Serena Villata, and Asuncion Gomez-Perez. 2014. A dataset of RDF licenses. In *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014 (Frontiers in Artificial Intelligence and Applications)*, Rinke Hoekstra (Ed.), Vol. 271. IOS Press, 187–188. https://doi.org/10.3233/978-1-61499-468-8-187

[19] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. 2014. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th Int. Semantic Web Conf. (Lecture Notes in Computer Science)*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A Knoblock, Denny Vrandecic, Paul T Groth, Natasha F Noy, Krzysztof Janowicz, and Carole A Goble (Eds.), Vol. 8796. Springer, 245–260. https://doi.org/10.1007/978-3-319-11964-9_16

Linking and Negotiating Uncertainty Theories Over Linked Data

Ahmed El Amine Djebri
Université Côte d'Azur, Inria, CNRS, I3S, France
djebri.emp@gmail.com

Andrea G.B. Tettamanzi
Université Côte d'Azur, Inria, CNRS, I3S, France
tettamanzi@univ-cotedazur.fr

Fabien Gandon
Inria, Université Côte d'Azur, CNRS, I3S, France
fabien.Gandon@inria.fr

ABSTRACT

There is no credibility insurance measure for the information provided by the Web. In most cases, information cannot be checked for accuracy. Semantic Web technologies aimed to give structure and sense to information published on the Web and to provide us with a machine-readable data format for interlinked data. However, Semantic Web standards do not offer the possibility to represent and attach uncertainty to such data in a way allowing the reasoning over the latter. Moreover, uncertainty is context-dependent and may be represented by multiple theories which apply different calculi. In this paper, we present a new vocabulary and a framework for handling generic uncertainty representation and reasoning. The meta-Uncertainty vocabulary offers a way to represent uncertainty theories and annotate Linked Data with uncertainty information. We provide the tools to represent uncertainty calculi linked to the previous theories using the *LDScript* function scripting language. Moreover, we describe the semantics of contexts in uncertainty reasoning with meta-uncertainty. We describe the mapping between RDF triples and their uncertainty information, and we demonstrate the effect on the query writing process in *Corese*. We discuss the translatability of uncertainty theories and, finally, the negotiation of an answer annotated with uncertainty information.

CCS CONCEPTS

• **Information systems → Uncertainty**; **Web data description languages**; **Ontologies**; • **Theory of computation → Incomplete, inconsistent, and uncertain databases**.

KEYWORDS

Uncertainty, Ontology, Negotiation, Linked Data, Semantic Web

ACM Reference Format:
Ahmed El Amine Djebri, Andrea G.B. Tettamanzi, and Fabien Gandon. 2019. Linking and Negotiating Uncertainty Theories Over Linked Data. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317074

1 INTRODUCTION

The Web is a massive source of information, and the emergence of Semantic Web [3] technologies allowed the transition from a document-driven view of the Web to a data-driven one. As no credibility insurance measure is taken in most cases, information on the Web cannot be checked for accuracy. In fact, many websites can post whatever they want including biased, subjective, incomplete or uncertain information [20]. In addition, several issues involving the modeling of uncertain Web Data and services are raised by Benslimane *et al.* [2] arguing that uncertainty should be modeled and measured in order to be associated with a piece of information or a service. Such a model prepares the ground for querying, searching, composing and harnessing uncertain data. Semantic Web technologies aimed to offer both structure and sense to the existing Web. However, Semantic Web standards do not offer the possibility to represent and attach uncertainty to such pieces of data while keeping standard semantics without further extension.

Various theories exist and can be used to model different aspects or types of uncertainty. Moreover, the choice of one theory among the others depends on the context and the application. Thus we argue that the bridge between uncertainty and the Semantic Web is crucial because of the dependence of many applications on uncertain data, the existence of many uncertainty theories, the need of interoperability and reusability of uncertain data.

First of all, the definition of uncertainty itself is challenging. It can be *epistemic*, i.e., stemming from our ignorance (incomplete knowledge, lack of a model) of an entity or process, or *ontic*, i.e., representing the inherent randomness of a phenomenon or system (a roulette, for instance). In addition, the border between these two types of uncertainty is somehow blurred and arbitrary, in that it depends on our point of view and on the level of abstraction of knowledge representation.

Multiple uncertainty theories are currently applied in Artificial Intelligence in order to reconcile the data, correct them, extrapolate or predict new values or simply assess the degree of uncertainty of information or error comparing to a dataset [14]. Probability theory is but one example, if not the best known and most time-honored, of such theories, which focuses on the representation and manipulation of the ontic uncertainty. Other theories are interested in the quantization (or "granulation") of vague data. As the words "tall" or "cold" do not have a specific measure, they still can be modeled using fuzzy-sets and manipulated using fuzzy logic. A non-exhaustive list of uncertainty theories and their formalization can be found in [17].

To integrate uncertainty in the Semantic Web stack, two aspects must be taken into account:

- **Syntax**: uncertainty needs to be represented in order to be queried or published. The format must be machine-readable and interchangeable. The vocabulary proposed by the W3C URW3-XG group [18] enables to annotate data with the type, the model and the derivation of uncertainty. The group offers a limited list of models (Fuzzy sets, rough sets, etc.) with which neither information regarding the quantification

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317074

of uncertainty nor the specificities of each approach and theory are provided.
- **Semantics**: reasoners are used on top of ontologies to infer new triples (based on OWL profiles) or validate the knowledge base with the schema. Direct semantics in OWL [8] is intended to give sense to ontology structures compatible with the $\mathcal{SROIQ}$ description logic. Faced with uncertainty as described below and assuming that triples can be annotated with uncertainty values, OWL does not provide the necessary tools to manipulate uncertainty, as, with the generation of new information, uncertainty must be provided too. Several extensions to the standards are done in the Schema layer itself such as FuzzyOWL [22] or PossOWL [19], while other addressed the 'Unifying logic' layer like FuzzyDL [11]. Such standards can be extended too to enable inferring new information about uncertainty or keep only triples that follow a defined set of rules in the knowledge base.

Dividino *et al.* [12] presented a framework extending both RDF and SPARQL to enable meta-knowledge querying. Their framework enables the system administrator to define meta-knowledge properties and for each, should define the intended semantics and the knowledge dimension inside the application in a non-standard format. Thus, neither semantics for the dimensions are publishable, nor semantics for uncertainty itself are provided.

Another problem that arises from the different theories is the relationship between theories and data. Some theories are suitable for specific contexts and applications. Which requires either finding the suitable theory to annotate the data, or a suitable transformation of the existing annotations to fit the requirement. Both reasons raise the question of alternatives and negotiation of content with the data sources, alongside with the possibility for users to represent their preferences in queries.

In this paper, we focus on a generic representation for uncertainty information, one that can be publishable and reusable. We summarize the *mUnc* vocabulary enabling both the representation of uncertainty theories with their calculi and the annotation of data with uncertainty information. We present an extension to RDF Semantics by giving a contextual meaning to Named Graphs. We offer the possibility to use *mUnc* alongside with our framework built on top of the *Corese* Semantic Web engine [10] and using the *LDScript* function scripting language [9] based on the SPARQL filter language, to map between triples and their uncertainty metadata, using multiple mapping modes. We discuss the translatability of uncertainty theories and propose to extend *mUnc* to fill that gap, and then discuss the negotiation of theories over HTTP as a special case of *conneg* (Content Negotiation) where clients can state their preferences in terms of uncertainty theories and servers can select or translate among theories to serve the best possible answer.

The rest of the article is organized as follows. Section 2 first recalls the uncertainty in information theory and discusses a model for the existing uncertainty theories in the literature. Then, Section 3 discusses the integration of uncertainty in the semantic web, offering a presentation of the *mUnc* vocabulary, discussing the publishing of uncertainty approaches and their calculi on the Semantic Web, and giving an overview of metadata mapping modes and how to query over uncertain Linked Data. Section 4 is dedicated to the translatability of uncertainty theories and the negotiation details. Related works are discussed in Section 5 while the last section summarizes our work and offers a glimpse over the perspectives.

2 UNCERTAINTY IN INFORMATION

Information uncertainty can present itself in different forms depending on the domain of the definition. Data itself can be wrong from the source: entry mistakes, deceit, ambiguity, sensor errors, etc. In the context of decision-making, we link uncertainty to the outcome by defining uncertain data as "entries leading to a wrong output". Whether the data are incomplete (or imprecise), vague (or fuzzy, ambiguous), incorrect (or invalid). We find the previous specifications under other definition in the literature. For example, Dubois *et al.* [14] state that incompleteness, uncertainty, graduality, and granularity are different. They consider uncertainty as a measure for ignorance of the truth of a primitive item of information (proposition, statement, a subset of possible values, etc.), and quantified by a numerical or symbolic token located in the metalevel. In their context, the authors discuss assigning to each primitive item (fine grain) of information A, a number $g(A) \in \{0, 1\}$ which evaluates the confidence of an agent in the truth of a proposition asserting $v \in A$. The presented property is also called *confidence function*, sometimes *capacity* [7] or *fuzzy measure* [23]. This measure preserves:

- monotonicity : If $A \subseteq B$ then $g(A) \leq g(B)$;
- $g(\bot) = 0, g(\top) = 1$;

The confidence function can have multiple types according to the data in measure. It can be a *possibility*, *necessity* or even a *probability* measure. The former two measures are dual and the latter satisfies the additivity property. Several theories were established in order to formalize uncertainty, such as probability theory, possibility theory, Dempster-Shafer evidence theory, belief functions, etc. Each of the previous has a set of measures and a logic to read or to infer new values. For instance, possibility theory states that given an event A, which is a set of interpretations or outcomes, the possibility of the event to occur is given by the possibility distribution:

$$\Pi(A) = \sup_{\mathcal{I} \in A} \pi(\mathcal{I})$$

While laying under the Open World Assumption, inconsistency is another challenging aspect of uncertain data where the existence of multiple interpretations of the same knowledge base leading to undecidability. As an alternative of two-valued logic, other types of logic are exploited: defeasible logic based on rules or paraconsistent logic are among those. Such alternatives offer to draw conclusions from partial and conflicting information.

Uncertainty is also affected by the context of the reading [12]. A distributive reading of uncertainty means that statements in a set provide each a portion of the uncertainty value of the whole set. Collective reading, on the other hand, assigns to the whole set an uncertainty measure. The two readings are not exhaustive, an alternative would be inheriting the uncertainty measure of the set.

Klir *et al.* [17] observe that dealing with uncertainty consists of four distinct levels: representation, calculus, measurement, and development of methodological aspects of the theory. Our focus in this paper is on the first two levels.

3 INTEGRATING UNCERTAINTY IN THE SEMANTIC WEB

Semantic Web [3] is meant for machines to understand and infer knowledge as humans do while exploring the Web. Granularizing and linking pieces of data allow machines to provide more relative content and help in the query resolution process. Schema and ontology languages offered a backbone to the existing information by providing the vocabulary of a defined topic, the relationships ruling over the different terms and the semantics allowing to logically establish links in between. Several standards (RDF, RDFS, OWL, SPARQL, etc.) of the Semantic Web technology stack enable the representation and querying of RDF data.

To integrate uncertainty in the Semantic Web, this dimension must comply with the standards and, like any other data on the web, must be reusable and publishable. In this section, we offer an ontology covering uncertainty representation, and we show the possibility of replicating the semantics of uncertainty using calculi represented in the *LDScript* function scripting language.

3.1 mUnc: a vocabulary for uncertainty theories

To enable uncertainty representation on the Semantic Web, we need to opt for an interchangeable format to write uncertainty theories. We propose *mUnc* [1], an RDFS ontology for uncertainty theories. *mUnc* (for *m*eta-*Unc*ertainty) enables publishing uncertainty information based on uncertainty theories. Figure 1 gives an overview of the core concepts and properties of *mUnc* .

We have adopted the definition of *sentence* and *world* proposed in the URW3-XG ontology. A *sentence* is an expression evaluating a truth value, while the *world* represents the context in which a *sentence* is stated. Still, unlike the previous definition, both sentences and worlds can be annotated with *meta* information. For instance, the sentence `ex:S1` representing the triple ⟨`ex:StefanoTacconi`, `dbo:height`, 188⟩ referring to the height of the football player is stated in the context of the French language chapter of DBpedia [4], assuming that the latter is consistent [5]. *Uncertainty* is considered a specialization of the general concept of *meta*. This simplifies the task if any future extensions for other types of metadata such as provenance or trust are proposed. We do not include the concept of *Agent*, as it can be included using other vocabularies like the W3C *PROV*[2] Ontology.

An uncertainty theory (*Uncertainty Approach*) is linked to a set of *features* and *operators*. The *features* are the metrics on which uncertainty theory is based to indicate the degree of truth, credibility, or likelihood of a sentence. Each *feature* links a *value* to the *uncertainty* information. The *operators* represent the logic to apply to the previous *values*, while the *operations* are the implemented calculus for such logic. Other concepts in the URW3-XG ontology like the type or the derivation of uncertainty can be represented as features of an uncertainty approach.

To illustrate the previous definitions, we can annotate the previous sentence using probability theory. It can be represented using only one feature: the probability value. We choose three logical operators to include with the definition: and, or, not. The example in the listing 1 shows how to assert that a sentence `ex:S1` is true with a probability of 0.7. For the sake of illustration, we use reification to attach an IRI to the previous sentence, although no preference about metadata representation methods is stated [16].

[1] http://ns.inria.fr/munc/
[2] http://www.w3.org/TR/prov-o/

```
ex:S1 a rdf:Statement;
      rdf:subject ex:StefanoTacconi;
      rdf:predicate dbo:height;
      rdf:object "188"^^xsd:decimal.

ex:Probability a munc:UncertaintyApproach;
  munc:hasUncertaintyFeature ex:probabilityValue;
  munc:hasUncertaintyOperator ex:and, ex:or, ex:not.

ex:probabilityValue a munc:uncertaintyFeature;
  rdfs:range xsd:decimal.

ex:S1 munc:hasMeta [a munc:Uncertainty;
  munc:hasUncertaintyApproach ex:Probability;
  ex:probabilityValue 0.7].
```

Listing 1: Representing uncertainty with *mUnc*

3.2 Uncertainty calculi

Semantic Web ontology languages focus on classification ontological knowledge and do not support the provision of procedural attachments or functions inside ontologies. Our model allows linking the features of uncertainty approaches to their proper calculi (arithmetic, logical, comparison, etc). To represent the calculi, we rely on the *LDScript* function definition language [9], a programming language whose objects are RDF entities. It is built on top of SPARQL and relies on the SPARQL filter expression language. *LDScript* as a language permits variable declaration, assignment, function call, return, etc. Using *LDScript*, we can define functions named with an IRI and one or several arguments that are variables in the SPARQL syntax. This enables defining uncertainty operations and linking them to uncertainty features.

To continue with the previous example, considering the fact that the sentence `ex:S1` is true with a probability of 0.7, and is stated in a context `ex:C1` where all contained facts are considered to be true with a probability of 0.9. The probability of a conjunction of two supposed independent events A and B is given by Equation (1):

$$P(A \wedge B) = P(A) \times P(B). \tag{1}$$

Such value can be calculated for the user using the function referenced by `ex:multProbability` and defined in *LDScript* as shown in the following example:

```
function ex:multProbability(?pA, ?pB){?pA * ?pB}
```

Therefore, binding the function `ex:multProbability(0.7, 0.9)` during a SPARQL query execution will return 0.63. The former definition of the probabilistic approach using *mUnc* can be enriched by linking the IRI of the function to the declared feature, simply by adding the triple:

```
ex:ProbabilityValue ex:and ex:multProbability.
```

As stated before, each function is considered as a resource, due to the IRI defining its name. We can store such functions in SPARQL files all over the Web, and access their code using their reference.

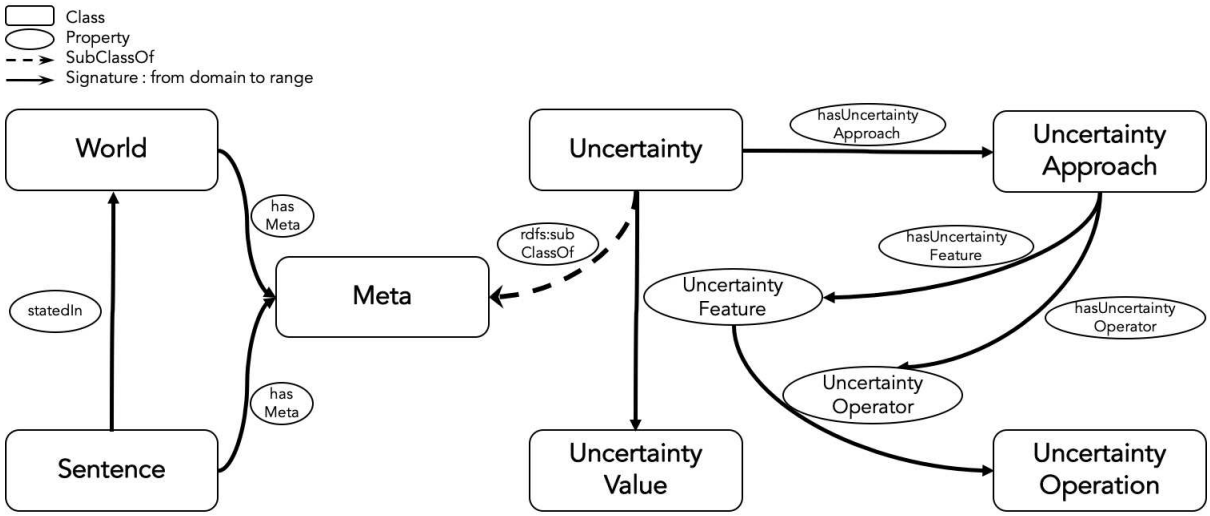


Figure 1: Overview of the *mUnc* ontology and its core concepts

3.3 Contextualizing, Mapping and Querying Uncertain Linked Data

As mentioned before, uncertainty depends on a context. A single world may accept different interpretations using different uncertainty measurements. Dubois [13] explains that beliefs are the different views upon a single world. We rely on the fact that data is issued by querying n uncertain data sources $s_1, s_2, ..., s_n$, each possibly containing several contexts C_{ij}, each representing consistent information. This means that each context contains a set of triples that do not lead to contradictory reasoning.

mUnc does not provide an extension of RDF Semantics. Instead, we rely on the SPARQL query language to provide a mapping between sentences and the uncertainty information presented to the user. Moreover, we consider *mUnc* as an approach to providing definitions of known and custom uncertainty theories, for which we do not provide any specific semantics. The possibility of defining a calculus alongside with the ontology is an alternative to generalize and to reuse of the shared rules between uncertainty theories such as maximizing or minimizing a feature.

We note $\mathcal{U}_{S_{C_{ij}}}$ the uncertainty information about the sentence S cited in the context C_{ij} and $\mathcal{U}_{C_{ij}}$ the uncertainty information about the context C_{ij}. Each sentence S stated in a context C_{ij} of a source s_i, will be mapped to a combined set of pairs (*Uncertainty Feature*, *Uncertainty Value*) issued from both sentence and context metadata (noted $\hat{\mathcal{U}}_{S_{C_{ij}}}$). This requires defining a metadata-mapping mode (see table 1).

The modes depend on the purpose of the application, the data itself, and the semantics of uncertainty theories. In the first mode, only uncertainty information linked to context C_{ij} is considered. The second mode considers only pairs from the lowest level of granularity, while the third mode enables inheriting context metadata but overrides the values for existing features in uncertainty information linked to the sentence.

In this paper, we use a specific meta mapping mode which relies on uncertainty calculus to evaluate a new set of pairs based on both information from sentences and contexts. In the *metaList* algorithm we implement and use the last mode in the previous table (see Listing 2),

```
@public
function munc:metaList(?xT, ?xG){
    let(SELECT ?xT ?xG (group_concat(?FV;separator="-") as ?metaD) WHERE
        {{
          SELECT ?xT ?xG (CONCAT(?xF,'=',?xV) AS ?FV) WHERE
          {
              ?xG ?xF ?xV1
              OPTIONAL {?xT ?xF ?xV2}
              ?xF rdfs:subPropertyOf munc:uncertaintyFeature
              ?xF ex:and ?xFFunction
              BIND(IF(BOUND(?xV2),funcall(?xFFunction,?xV1,?xV2),?xV1) AS ?xV)
          }
        } GROUP BY ?xT ?xG}
        UNION
        {{
          SELECT ?xT ?xG (CONCAT(?xF,'=',?xV) AS ?FV) WHERE
          {
              ?xT ?xF ?xV
              ?xF rdfs:subPropertyOf munc:uncertaintyFeature
              FILTER NOT EXIST {?xG ?xF ?xV2}
          }
        } GROUP BY ?xT ?xG}
        )
        {?metaD}
}
```

Listing 2: *metaList* algorithm in *LDScript*

The `munc:metaList` function is declared in the example as "`@public`". This keyword is implemented in *Corese* as many others (`@define`, `@visitor`, `@trace`, ...) defining specific routines in the Semantic Web engine. The former keyword allows the previous code to be accessed globally in the engine through its reference, and without the need to rewrite the function with each query. The listing 2 translates the `metaList` algorithm into *LDScript*. The result of binding this function in a SPARQL query is a string that groups all uncertainty features and their corresponding values from the Universal Uncertainty Information set of the corresponding sentence.

Table 1: Metadata mapping modes

Considered level of granularity	Mode	Explanation
Context	Imposed	$\hat{\mathcal{U}}_{S_{C_{ij}}} = \mathcal{U}_{C_{ij}}$
Lowest granularity only	Overriding	$\hat{\mathcal{U}}_{S_{C_{ij}}} = \begin{cases} \mathcal{U}_{S_{C_{ij}}} & \text{if } \mathcal{U}_{S_{C_{ij}}} \neq \varnothing \\ \mathcal{U}_{C_{ij}} & \text{otherwise} \end{cases}$
Lowest granularity first	Override if Exists	$\hat{\mathcal{U}}_{S_{C_{ij}}} = \mathcal{U}_{S_{C_{ij}}} \cup \{(F,V) \mid (F,V) \in \mathcal{U}_{C_{ij}}, \nexists V' \ (F,V') \in \mathcal{U}_{S_{C_{ij}}}\}$
Multi-level	Uncertainty Calculus	$\hat{\mathcal{U}}_{S_{C_{ij}}} = \{(F,V) \mid (F,V) \in \mathcal{U}_{C_{ij}}, \nexists V' \ (F,V') \in \mathcal{U}_{S_{C_{ij}}}\}$ $\cup\{(F,V) \mid (F,V) \in \mathcal{U}_{S_{C_{ij}}}, \nexists V'' \ (F,V'') \in \mathcal{U}_{C_{ij}}\}$ $\cup\{(F,V) \mid \exists V_{C_{ij}}, V_{S_{C_{ij}}} \ (F,V_{C_{ij}}) \in \mathcal{U}_{C_{ij}}, (F,V_{S_{C_{ij}}}) \in \mathcal{U}_{S_{C_{ij}}},$ $V = eval(Calculus(F), V_{C_{ij}}, V_{S_{C_{ij}}})\}$

Corese also implements *Linked Functions* enabling storing *LDScript* [9] functions in external SPARQL query files on the Web. Such functions, referenced by IRIs, may be called at the moment of query execution. The former feature permits publishing and executing the calculi of uncertainty approaches. Additionally, this feature may be extended to allow the capitalization of existing software libraries from other programming languages like C++ or Java.

The Semantic Web engine also allows defining specific routines preceding the query execution. One can integrate query transformation or precalculations of some variables. We implemented the previous meta-mapping mode in extension to the visitor `"@metadata"` and enabled rewriting SPARQL queries to simplify querying for uncertainty information. Using `"@metadata"` and with `munc:metaList` publicly defined, querying for the height of the football player Stefano Tacconi in a data source is as follows.

```
@metadata
   prefix ex: <http://example.org/> .
   prefix munc: <http://ns.inria.fr/munc/> .

   SELECT * WHERE {
     ex:StefanoTacconi dbo:height ?o
   }
```

Listing 3: Query rewriting using visitors implementing the access to uncertainty information

4 NEGOTIATING UNCERTAINTY ON THE SEMANTIC WEB

In addition to the previous two-step process leading to the generation of Universal Uncertainty Information Sets alongside with sentences, users may actually have a preference for one theory or another.In this section, we will discuss the translatability between uncertainty theories and how, using HTTP content negotiation (conneg), users may negotiate the theory they want for their results.

4.1 Translating uncertainty between theories

Many examples reject the claim that uncertainty can be represented only using probability theory. However, the belief about uncertainty being the lack of information or the deficiencies due to a shortage of knowledge urge researchers to believe that it may be unified, or at least, that the different views may be linked. Dubois *et al.* [15] stated that transformation is useful in any problem where heterogeneous uncertain and imprecise data must be dealt with (e.g. subjective, linguistic-like evaluations and statistical data). Zadeh [25] cites the example of Dempster–Shafer theory which is a theory of random sets. The latter are a probability distribution of possibility distributions. An interesting analysis of the possibility-probability transformation and its links to graphical models can be found in [1].

With the use of our framework, each and every context will issue an answer to the user. If the answers are annotated with the same theory and the same set of features, this enables ranking the results or offers more options to control the results. In the example of search engines, this could support a uniform criteria to order the results shown to the user. However, on an open Web where several open sources are queried, the results might use different features from different theories.

A translation must offer to transform a Universal Uncertainty Information Set $\hat{\mathcal{U}}_{S_{C_{ij}}}$ of a sentence S annotated following an uncertainty approach T_1, to another set annotated with a different uncertainty approach T_2. The translatability of theories should take into account several issues such as the symmetry, the reversibility, and the possible loss of information.

To fit in with the previous requirements, we define a translatability relationship between two uncertainty theories as follows:

Definition 4.1. A theory T_1 has a *translatability* relationship with a theory T_2, if there exists a mapping $M : F_{T_1} \rightarrow F_{T_2}$ from the set of features F_{T_1} represented in theory T_1 to the set of features F_{T_2} represented in theory T_2 such that every possible feature of F_{T_1} is mapped to a set of feature of F_{T_2} semantically coherent with the uncertainty initially expressed in T_1. We note $T_1 >| \ T_2$.

The former definition is valid for all theories that have a relationship allowing the conversion of features from one theory to another, regardless the loss of information. In case the conversion does not generate a loss of information allowing the reversibility of the operation, we define the relationship as follows:

Definition 4.2. A theory T_1 has an *ideal translatability* relationship with a theory T_2, if T_1 is translatable to T_2 ($T_1 >| \ T_2$) and there is no loss of information in the translation. We note $T_1 \gg T_2$

We should mention that an ideal translatability might not be reversible, regardless the semantics of the translatability. The loss of information disables the backward operation. If the other case

is considered, where we have no loss of information, then we can define a full translation as follows:

Definition 4.3. A theory T_1 has a *full translatability* relationship with a theory T_2, iff T_1 ideally translatable to T_2 ($T_1 \gg T_2$) and, inversely, T_2 ideally translatable to T_1 ($T_2 \gg T_1$). We note $T_1 \otimes T_2$.

Using our *mUnc* vocabulary and the framework previously proposed, we are able to formalize the translation (if it exists) between the different theories. For this, we extended *mUnc* with the set of the following properties:

- `munc:hasTranslation` (definition 4.1)
- `munc:hasIdealTranslation` (definition 4.2)
- `munc:hasFullTranslation` (definition 4.3)

These QNnames respectively identify the previous definitions. Figure 2 shows the extension, where each property of the previous set has for domain an uncertainty approach, and for range a blank node pointing to both the destined theory and the IRI of the translation function, written in *LDScript*.

By definition, the translativity properties have a number of algebraic properties:

Lemma 4.4. *(transitivity of translatability) Let $T_i, i = 1, 2, 3$ be three uncertainty theories. If $T_1 >| T_2$ and $T_2 >| T_3$ then $T_1 >| T_3$.*

Lemma 4.5. *(reflexivity of translatability) Each uncertainty theory has a full translatability with itself. We note $T \otimes T$.*

Lemma 4.6. *(Symmetry of full translatability) Let T_1, T_2 be two uncertainty theories. If $T_1 \otimes T_2$ then $T_2 \otimes T_1$.*

Lemma 4.7. *(full translatability is an equivalence relation) because it is transitives, symmetric and reflexive.*

We can note that full translatability being an equivalence relation it allows to form equivalence classes by transitive closure i which we have translatability with no loss of information from a theory T_i to any other theory T_j of its class. We note this set $TC_{\mathcal{U}}(T_i)$.

To illustrate the previous extension, we propose to represent the example proposed in [21] about the Optimal Transformation (OT) from probability to possibility. We declare a translatability relationship between probability theory `ex:Probability` and possibility theory `ex:Possibility` representing the two different uncertainty theories. We enrich the data source with the triples below, where `ex:translateProbaToPoss` is an *LDScript* function.

```
ex:Probability munc:hasTranslation ex:Possibility.

ex:translateProbaToPoss munc:translateFrom ex:Probability.
ex:translateProbaToPoss munc:translateTo ex:Possibility.
```

4.2 Negotiation of Uncertainty Headers

Based on the previous model we can now support the possibility of negotiating answers annotated with different uncertainty theories. Content negotiation can be based on HTTP headers or non-HTTP methods such as query arguments in IRIs. Following the W3C working draft proposed by Svensson *et al.* [24] we propose that clients may negotiate a representation annotated with a specific uncertainty theory, using q-values to express their preference regarding the uncertainty theories they are to receive. Since uncertainty theories are already defined using *mUnc* and named with IRIs, both server and client can exchange and verify the conformity of their options. We propose to handle three use cases:

(1) Uncertainty information exists in the queried source in one or many requested uncertainty theories. We answer with the first theory selected by the user. In the example, uncertainty information is issued from a context annotated with `evidence`.

```
GET /some/resource HTTP/1.1
Accept: text/turtle;q=0.9;uncertain="http://example.org/probability",
       text/turtle;q=0.7;uncertain="http://example.org/evidence";

HTTP/1.1 200 OK
Content-Type: text/turtle;uncertain=<http://example.org/evidence>
```

(2) The data source does not offer information about all requested theories, but a translation from existing uncertainty information to one or more requested theories is available. In this example, uncertainty is available in probability theory. The returned information are evaluated using the function `ex:translateProbaToPoss` and presented to the client with an indication about the type of translation the data underwent.

```
GET /some/resource HTTP/1.1
Accept: text/turtle;q=0.7;uncertain="http://example.org/possibility";

HTTP/1.1 200 OK
Content-Type: text/turtle;uncertain=<http://example.org/possibility>;
       translation=full
```

We note that the selection of a suitable translation starts from the transitive closure of full translations $TC_{\mathcal{U}}$(`possibility`) offering more information and graduates to the normal translatability relationship.

(3) The data source has no information about the theory and no available translations, we answer the user with the existing information. The default uncertainty information proposed by the server is returned in such case.

```
GET /some/resource HTTP/1.1
Accept: text/turtle;q=0.9;uncertain="http://example.org/probability",
       text/turtle;q=0.7;uncertain="http://example.org/evidence";

HTTP/1.1 200 OK
Content-Type: text/turtle;uncertain=<http://example.org/possibility>;
       default=true
```

5 RELATED WORKS

mUnc can extend the work done by Cabrio *et al.* [6] by enriching the proposed fuzzy labeling algorithms with definitions of other uncertainty theories that can be more suitable to the data. The Linked Data sources can adopt this approach to enrich federated queries with uncertainty information and, progressively, build a consensus-based Linked Data source. A set of other applications such as fake news detection (definition of a theory and logic for fake news), argumentation-based systems and even community-based data sources such as DBpedia can use *mUnc* to enrich their future content with uncertainty information. Furthermore, *mUnc* shares the same objectives as the W3C Credible Web community group [20] as for exchanging data which bears directly on credibility assessment while keeping standardization in data interchange.

The focus on uncertainty translatability was mainly in AI-based applications. We point that Semantic Web requirements in term of interoperability and information usage are very important and

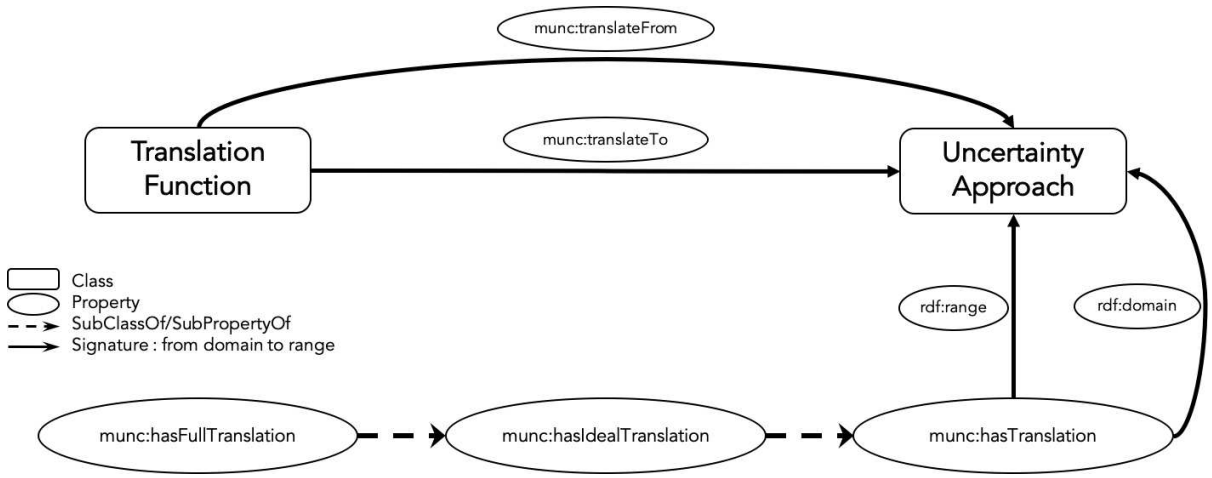


Figure 2: Extending *mUnc* ontology with translation properties

the suitability of uncertainty theories to different types of data and applications need to be further discussed.

6 CONCLUSION AND PERSPECTIVES

In this paper, we discussed the representation and publication of uncertainty on the Semantic Web. We presented a vocabulary allowing the representation of uncertainty theories and the annotation of sentences using the Semantic Web standards. We explained the publishing of reusable uncertainty calculus using *LDScript*. We also offered the possibility to translate between uncertainty theories and to negotiate uncertainty information following a specific theory.

Uncertainty representation is the first step of a long process, including the preliminary calculus of uncertainty values and the propagation of uncertainty among interconnected Linked Data sources. The translation process is also a first step enabling, to some extent, merging uncertain data annotated using different uncertainty approaches. In our future work, we would like to implement context overlapping, allowing the selectivity inside the source between contexts and the optimization of the storage. We would also study the relationship between data, applications, and uncertainty theories that are used.

ACKNOWLEDGMENTS

We thank Olivier Corby for his insights about *Corese* and *LDScript*. This work is supported by the Algerian government.

REFERENCES

[1] Salem Benferhat, Amélie Levray, and Karim Tabia. 2015. On the analysis of probability-possibility transformations: Changing operations and graphical models. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, 279–289.
[2] Djamal Benslimane, Quan Z Sheng, Mahmoud Barhamgi, and Henri Prade. 2016. The uncertain Web: Concepts, challenges, and current solutions. *ACM Transactions on Internet Technology (TOIT)* 16, 1 (2016), 1.
[3] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284, 5 (2001), 34–43.
[4] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia-A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the world wide web* 7, 3 (2009), 154–165.
[5] Elena Cabrio, Serena Villata, and Fabien Gandon. 2014. Classifying Inconsistencies in DBpedia Language Specific Chapters.. In *LREC*. 1443–1450.
[6] Elena Cabrio, Serena Villata, and Alessio Palmero Aprosio. 2017. A RADAR for information reconciliation in Question Answering systems over Linked Data 1. *Semantic Web* 8, 4 (2017), 601–617.
[7] Gustave Choquet. 1953. Theory of capacities. *Annales de l'institut Fourier* 5 (1953), 131–295. https://doi.org/10.5802/aif.53
[8] World Wide Web Consortium et al. 2012. OWL 2 web ontology language document overview. (2012).
[9] Olivier Corby, Catherine Faron-Zucker, and Fabien Gandon. 2017. LDScript: a Linked Data Script Language. In *International Semantic Web Conference*. Springer, 208–224.
[10] Olivier Corby and Catherine Faron Zucker. 2002. Corese: A corporate semantic web engine. In *International Workshop on Real World RDF and Semantic Web Applications, International World Wide Web Conference*.
[11] Valerie Cross and Shangye Chen. 2018. Fuzzy Ontologies: State of the Art Revisited. In *Fuzzy Information Processing*, Guilherme A. Barreto and Ricardo Coelho (Eds.). Springer International Publishing, Cham, 230–242.
[12] Renata Dividino, Sergej Sizov, Steffen Staab, and Bernhard Schueler. 2009. Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *Web Semantics: Science, Services and Agents on the World Wide Web* 7, 3 (2009), 204–219.
[13] Didier Dubois. 2011. Information Fusion and Revision in Qualitative and Quantitative Settings. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, 1–18.
[14] Didier Dubois and Henri Prade. 2010. Formal Representations of Uncertainty. In *Decision-making Process: Concepts and Methods*. 85–156. https://doi.org/10.1002/9780470611876.ch3
[15] Didier Dubois, Henri Prade, and Sandra Sandri. 1993. On possibility/probability transformations. In *Fuzzy logic*. Springer, 103–112.
[16] Johannes Frey, Kay Müller, Sebastian Hellmann, Erhard Rahm, and Maria-Esther Vidal. 2017. Evaluation of Metadata Representations in RDF stores. *Semantic Web Journal* (2017).
[17] George J Klir and Richard M Smith. 2001. On measuring uncertainty and uncertainty-based information: recent developments. *Annals of Mathematics and Artificial Intelligence* 32, 1-4 (2001), 5–33.
[18] Kenneth J Laskey and Kathryn B Laskey. 2008. Uncertainty Reasoning for the World Wide Web: Report on the URW3-XG Incubator Group.. In *URSW*. Citeseer.
[19] Bal-Bourai Safia and Mokhtari Aicha. 2014. Poss-OWL 2: Possibilistic Extension of OWL 2 for an uncertain geographic ontology. *Procedia Computer Science* 35 (2014), 407–416.
[20] Sandro Hawke. 2018. Technological Approaches to Improving Credibility Assessment on the Web. https://credweb.org/report/20181011. Accessed: 2019-01-20.
[21] Yosra Ben Slimen, Raouia Ayachi, and Nahla Ben Amor. 2013. Probability-possibility transformation. In *International Workshop on Fuzzy Logic and Applications*. Springer, 122–130.
[22] Giorgos Stoilos, Giorgos B Stamou, Vassilis Tzouvaras, Jeff Z Pan, and Ian Horrocks. 2005. Fuzzy OWL: Uncertainty and the Semantic Web.. In *OWLED*.
[23] Michio Sugeno. 1977. Fuzzy measures and fuzzy integrals: A survey. 78 (01 1977).
[24] Svensson, Lars G. and Atkinson, Rob and Car, Nicholas J. 2018. *HTML 5.3*. W3C Working Draft. W3C. https://www.w3.org/TR/dx-prof-conneg/.
[25] Lotfi A Zadeh. 2011. Generalized Theory of Uncertainty: Principal Concepts and Ideas. In *Fundamental Uncertainty*. Springer, 104–150.

© Companion Proceedings of the 2019 World Wide Web Conference, published under Creative Commons CC-BY 4.0 License. DOI - https://doi.org/10.1145/3308560.3317076

The FAIR TRADE Framework for Assessing Decentralised Data Solutions

John Domingue

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, john.domingue@open.-ac.uk

Allan Third

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, allan.third@open.ac.uk

Manoharan Ramachandran

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, manoharan.ramachand-ran@open.ac.uk

ABSTRACT

Decentralised data solutions bring their own sets of capabilities, requirements and issues not necessarily present in centralised solutions. In order to compare the properties of different approaches or tools for management of decentralised data, it is important to have a common evaluation framework. We present a set of dimensions relevant to data management in decentralised contexts and use them to define principles extending the FAIR framework, initially developed for open research data. By characterising a range of different data solutions or approaches by how TRusted, Autonomous, Distributed and dEcentralised, in addition to how Findable, Accessible, Interoperable and Reusable, they are, we show that our FAIR TRADE framework is useful for describing and evaluating the management of decentralised data solutions, and aim to contribute to the development of best practice in a developing field.

CCS CONCEPTS

• Information systems~Distributed storage • Information systems~Resource Description Framework (RDF) • Security and privacy~Trust frameworks • Computing methodologies~Model verification and validation

KEYWORDS

Distributed Data, Evaluation, Open Data, Semantic Blockchain

1 Introduction

We now live in a world where a small number of large technology companies hold a significant amount of personal data. For example, the Guardian journalist Dylan Curran (Curran 2018) found that Google had over 5.5GB of personal data on him including everywhere he had been, and all of his Google and YouTube searches (including deleted searches). His analogous figure for Facebook was over 600MB of data. As highlighted in a number of recent scandals in the press, such as those associated with the misuse of data by Facebook and the now defunct company Cambridge Analytica (Guardian 2015), over-centralisation can cause problems for individuals as well as negatively affect democracy and national culture. The current situation has resulted in, according to the Web's inventor, "producing—with no deliberate action of the people who designed the platform—a large-scale emergent phenomenon which is antihuman." (Lee 2018a). There is also now a growing recognition of the social problems caused when centralised data systems are impoverished with respect to disadvantaged citizens, who are in effect 'data poor'. A report by the US Government's Consumer Financial Protection Bureau (Bureau 2015) found that 26 million US consumers (11% of the adult population) are 'credit invisible' and a further 19 million (8.3% of the adult population)

had credit data that was unscorable. In the UK, a lack of citizenship data led to 63 British citizens being deported (Grierson 2018) and up to 57,000 UK citizens could lose rights to homes, jobs, social benefits, UK National Health Service treatment, or be threatened with deportation in the near future (Fact 2018) – a situation which has been blamed on the combination of poorly-managed centralized data and a lack of official recognition of citizen-held data.

One type of approach to resolving the above problem has been through decentralising the data. In particular, replacing central data controllers with either the user herself, or with peer communities. Giving users control over their data and how it is processed is seen by a growing number of researchers and developers as the best mechanism to empower users and give us the 'internet we want'. The best known of these is the distributed ledger (Walport 2015) an architecture for maintaining a peer network of verified data which underpins the Bitcoin cryptocurrency. Within the Linked Data (LD) arena, Solid (Lee 2018b) provides users with their own data pods, and Linked Data Fragments (LDF) (Verborgh et al. 2016) enables data processing to be split between servers and clients. We have begun exploring the use of combined blockchain/Linked Data architectures within our own LinkChains work (Third and Domingue 2017).

Decentralisation has the potential to address problems such as those described above, but only when implemented in accordance with sound principles of decentralised data management. The problem we see for the research community is that there is currently no way of comparing decentralised data solutions as they vary greatly and, as far as we know, no comparison framework or standard exists. Our aim in this paper is to address this issue by presenting our FAIR TRADE framework for assessing decentralised data solutions, based on relevant dimensions for decentralization, and clear principles for best practice. Our framework builds upon the FAIR data principles (Wilkinson et al. 2016) which outline how scientific data should be managed, and adds dimensions related to decentralised data. Note that, just as the FAIR principles are focused on the *management* of data, so too are the TRADE principles we outline. More detailed dimensions of evaluation relating to *use* of data systems, such as performance, are not considered here.

The rest of this paper is organised as follows. In section 2 Background and Related Work, we first discuss the meanings of key terms and describe the FAIR data principles, and follow by describing a number of decentralised and distributed approaches to data management including blockchains, Solid and Linked Data Fragments, and our own decentralised approach to data management, LinkChains. Section 3 then describes our FAIR TRADE framework for assessing decentralised data approaches, before in section 5 testing our framework through the evaluation of six different data platforms. In the final section, we summarise and conclude our contributions.

2 Background and Related Work

2.1 Decentralisation and Distribution

There are a number of different ways in which data can be "spread out" across multiple locations. Each location could contain full copies of the same dataset, or different locations may contain different data. Independently, data at each location may be controlled or coordinated by a single central source, or different locations may be independent. In this paper, we refer to the former as the *distribution* of data, and to the latter as *decentralization* of data. The general theme in the literature appears to be loosely that distribution refers to location and decentralization to independence, but there is no clear consensus on definitions – e.g., Asano et al. (Asano et al. 2018) define decentralization to be when "data are maintained in different sites with autonomous storage and computation capabilities", which would make decentralised data systems a subset of distributed ones as defined by, e.g., Öszu & Valduriez (Özsu and Valduriez 2011) ("a number of autonomous processing elements […] interconnected by a computer network and that cooperate in performing their assigned tasks"). Interpreting these terms independently in the way we do permits greater expressive power, giving the option to describe all four combinations of decentralised or not vs. distributed or not.

Motivations for data distribution include redundancy, to prolong the lifespan of data and preserve it from accidental loss or malevolent attacks, as well as potentially spreading the load of data access beyond a single point of failure. Distribution of data need not be limited to simply storage and copying, but could, for example, also include querying: queries can be computationally expensive, and a solution including distributed query evaluation has the potential for better performance or sustainability than with non-distributed querying. Motivations for decentralised data systems are varied, whether a political wish to increase user control, or a practical need to use independent datasets spread across a network (e.g., the federated Linked Data model).

2.2 Autonomy and Identity

When the control of data is not restricted to those who have control over those locations in which it is stored (and, indeed, restricted *from* them), and is instead kept in the hands of particular individuals, e.g., the data owner, a data solution can be said to support *autonomy* of data. This is independent of decentralization and distribution; a fully centralized and non-distributed data system may nonetheless support autonomy if, for example, all data and associated metadata were to be encrypted under the data owners' control. There are a number of aspects of control which could be considered under data autonomy. Control over read/append/modify actions means that an individual can decide how much, if any, of their data is exposed, shared, added to, or changed, for how long and for what purpose. Particularly with personal data, control over usage and analysis may limit some of the recent abuses of data which we highlighted above. The format, metadata, and distribution of a dataset can also be controlled. Autonomy in this sense is an individual-focused concept, relative to particular users. As such, *identity* of users is important to analyzing it.

2.3 Trust

Trust in general, of course, is a broad term and includes aspects which are very difficult or impossible to address in technological systems. For example, a system cannot guarantee that data (e.g., about educational qualifications) was not fabricated prior to publication. Other aspects of trust are also covered by other principles or concepts. For example, trust that data is solely under the appropriate individual or institutional control falls under autonomy; integrity in the sense of conformance to a given schema falls under the Reusable principle of the FAIR standard, and so on. But there are some specific aspects of trust independent of those principles, and which specifically relate to decentralised data – in particular, properties relating to provenance metadata and integrity in the sense of content persistence over time. Specifically, what degree of assurance can a system provide that the publisher of a piece of data and the context of publication (e.g., date and time) can be correctly identified, and that claims based on the data are accurately based in the data originally published? In the centralised case, these are often guaranteed by the central data store, and trust in these aspects reduces to trust in the central store or its owners. Without a regulatable central authority to appeal to, these are important aspects to consider independently. (Schneider and Trustworthiness 1999) provides a high-level overview of trust in networked environments.

2.4 FAIR Data Principles

The FAIR principles emerged from a workshop held in Leiden in 2014 in recognition of the fact that the scientific community had not paid enough attention to the way digital objects including data are managed (Wilkinson et al. 2016). In particular, a need was recognised to improve the infrastructure associated with the use and reuse of scholarly data. A community of researchers, librarians, publishers and funders setup a group FORCE11 (Future of Research Communications and e-Scholarship) which agreed a set of minimal principles for the stewardship of digital scholarly artifacts. The resulting four FAIR principles are:

- ***Findable*** - data given globally unique persistent identifiers and are described via rich metadata. Metadata also has unique persistent identifiers.
- ***Accessible*** - identifiers can be used to retrieve data and metadata via a standard communications protocol which includes authentication and authorization when necessary.

- ***Interoperable*** **-** data and metadata are represented in a formal, accessible and applicable knowledge representation language. Vocabularies are also used to enable interoperability for both data and metadata. Qualified references are contained in (meta)data to point to other (meta)data.
- ***Reusable*** - data and metadata are described with relevant attributes, clear licenses and detailed provenance information.

The above principles have proved popular with the scientific community and have spawned a number of initiatives such as GO FAIR[1] which links the principles to the European Open Science Cloud. The FORCE11 community now has over 2,600 active members.[2] The FAIR principles are not inherently specific to scientific data management, and, as Mons, et al., (Mons et al. 2017) discuss, have begun to be adopted beyond science.

Wilkinson et. al. (Wilkinson et al. 2017) recently discussed a number of additions to FAIR including privacy protection via containers (based on Linked Data Platform containers)[3] and MetaRecords which return metadata given an HTTP GET request.

3 The FAIR TRADE Framework

Evaluation frameworks such as FAIR apply to both centralised and decentralised data solutions with little modification, but, in the decentralised case, we argue that they are insufficient. There are aspects of decentralisation which can have a significant effect on possible use of data, and which should be accounted for in any evaluation framework. To address this, we propose the FAIR TRADE framework for assessing decentralised data solutions. FAIR TRADE is an extension of FAIR: findability, accessibility, interoperability and reusability remain relevant dimensions for evaluation. In addition, we add dimensions of being TRusted, Autonomous, and Distributed and dEcentralised to form TRADE. Assessing a data solution according to FAIR TRADE provides a clear characterisation of its properties in terms of decentralisation and reusability.

While there is no essential need to modify the existing FAIR principles to apply to decentralised data, there are particular aspects worth noting that may require attention in interpreting them. For example, appropriate adaptations to ensure global uniqueness and persistence of identifiers may need to be made in a decentralised scenario to support Findability. The authorisation and authentication are harder in decentralized autonomous contexts, so they need more attention. The principles of Interoperability and Reusability are more straightforwardly agnostic of (de)centralisation status.

Our contribution

Decentralisation brings with it a number of further dimensions along which data management can vary, and which relate to the quality of a data source. We propose that the dimensions of trust, autonomy, distribution and decentralisation are effective aspects of data management to consider in decentralised contexts, and describing where in a FAIR TRADE space a data source lies is a useful characterisation of its nature, with adoption of *all* of the principles an indicator of best-practice in decentralised data stewardship.

1. ***TRusted*** – *T1: Data publication metadata can be automatically verified and validated* (e.g., that the publishing individual or organization, and timestamp of publication can be checked).

T2: *Claims made on the basis of data contents can be automatically verified* (e.g., that data describes the award of an educational certificate to the data owner). In interpreting this principle in specific contexts, the concepts and mechanisms of Verifiable Claims (as per the W3C Working Group) are useful: trustable access to properties of data, which is privacy-respecting, cryptographically secure and automatically verifiable.

[1] https://www.go-fair.org

[2] https://www.force11.org/community/members-directory

[3] https://www.w3.org/TR/ldp/#ldpc-container

Both of these principles are aided by support for a robust notion of identity, with attestations. This is important for data attribution and claim interpretation, and the various Verifiable Claim roles of holder, issuer, inspector-verifier and identity registry can all be relevant.

2. **Autonomous** – AU1: The owner(s), or authorized controller(s), of a piece of data have control over access to, and use of, data.

AU2: Data relating to personal identity follows AU1 and is trusted in the sense of T2.

Note that autonomy of data is not the same as accessibility under the FAIR principles. One may have access to, and have the credentials to have access to, one's data without having autonomous control over it. For example, personal data stored on the Facebook platform can be accessed by authorised users, but those users, including the person whose data it is, are not in control of it. In general, the principle of Autonomy relates to the legal and moral rights of data subjects over their own data.

3. **Distributed** – D1: In a given network, data is physically stored across some proportion or selection of nodes in that network.

D2: In a given network, the evaluation of data processing tasks can be executed across some specifiable proportion or selection of nodes.

Distribution is not necessarily a simple yes/no question: there are different degrees to which data, and data processing, can be distributed. Analysis of a system according to this principle supports, as appropriate, the characterization of quantified degrees of distribution for both D1 and D2 instead of a binary choice.

4. **dEcentralised** – E1: In a given network, no single node or small set of nodes controls which subset of data contents are held by any node or small set of nodes other than itself.

E2: In a given network, no single node or small set of nodes decides which new nodes can join.

A range of scenarios implementing the Trusted, Distributed and Decentralised principles are outlined in (Third and Domingue 2017), which we have since developed further into the LinkChains approach described in section 5.

4 Examples of Decentralisation

4.1 Distributed Ledges and Distributed Data Storage

It is important to distinguish between the terms 'distributed ledgers' and 'blockchains', which are often incorrectly used as synonyms. Distributed ledgers are replicated, shared and synchronised digital data geographically dispersed over multiple sites possibly including multiple institutions. A peer-to-peer network is required for communication and a consensus algorithm ensures replication and synchronisation across the multiple nodes. A key difference between applications that run on standard platforms and those that run on top of distributed ledgers is the way that data is stored and managed. Rather than connecting from a device (e.g. a mobile phone) to a central server, which holds all the required data (including private data), every player or volunteer in the network participates in ensuring that the whole network contains multiple copies of all the data. This changes a fundamental dynamic. The notion of centralised control disappears completely, rather data and computation are evenly owned, controlled and shared across the peer network. A blockchain is a specific type of distributed ledger where an ever-growing list of records, called blocks, are linked together to form a chain – hence the term 'blockchain'. The first blockchain was conceived by Satoshi Nakamoto in his white paper (Nakamoto 2008) as the basis for Bitcoin, the most famous blockchain based crypto-currency. The main idea behind Bitcoin was to create a currency specifically for the Internet rather than (as is the case in all fiat currencies) mapping an originally physical currency to the global communications infrastructure. Research produced by the University of Cambridge estimates that in 2017, there were 2.9 to 5.8 million unique users using a cryptocurrency wallet, most of them using Bitcoin.[4] The first issue that arises with Internet-based currencies is what is called the 'double spending problem'. This is the case when a digital 'coin' is spent, by an individual, for some service or good, and then the same coin is spent again by the same individual, for example, by copying or duplicating

[4] https://www.jbs.cam.ac.uk/faculty-research/centres/alternative- finance/publications/global-cryptocurrency

the relevant data. Blockchains address this problem by providing an immutable public ledger of all historical transactions. Once processed and stored within a block, and sufficient subsequent blocks have been confirmed, a transaction cannot be altered even by the transaction owners. Figure 1 shows a blockchain containing three blocks. Starting from the right, the newest block, each block points to its predecessor using a hash function. Transactions are stored in a Merkle Tree (Merkle 1980) - a tree of hashes - where the leaf nodes contain the transactions. This structure allows for efficient retrieval and ensures the veracity of the individual transactions in addition to the block – if a transaction is altered, then the hash will no longer be valid.

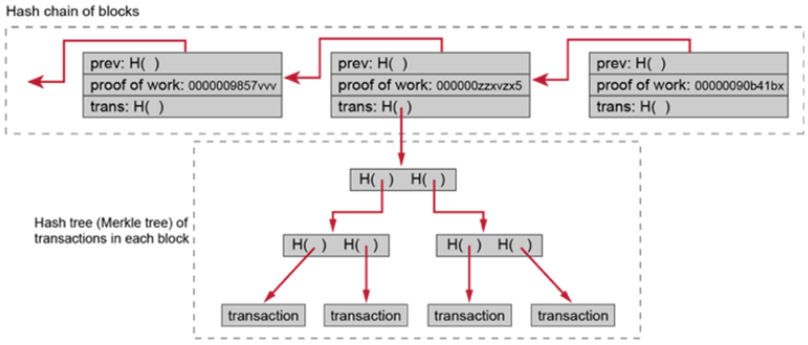


Figure 1: The hash links, proof-of-work nonce and Merkle tree of transaction data within a blockchain

The proof of work consensus mechanism which involves solving a cryptographic puzzle before anyone else has led to the growth of the computing power and electrical consumption associated with blockchain networks. Estimates are that by 2020 the Bitcoin network will expend as much electricity as Denmark.[5]b This has led to several blockchain platforms exploring other consensus mechanisms such as:

- ***Proof of stake*** – where the chances of being selected to produce the next block depend on the value of a 'stake' stored by a miner in a specific location. Variants of this take into account the 'age' of the stake.
- ***Proof of capacity*** – rather than the chances of being selected being related to the amount of computing power, as for proof of work, here the probability is related to the amount of storage a miner holds.
- ***Proof of burn*** – sending coins to an irretrievable address ('burn') gives one the right to be selected. The chances of being selected to mine the next block are related to the value of the burn.
- ***Proof of elapsed time*** – Intel has produced a special processor capability to implement a mechanism which relates elapsed time to the probability of being selected.

After Bitcoin, Ethereum (Buterin 2013) is the best known blockchain platform. Rather than serving as a platform for a cryptocurrency the underlying aim for Ethereum is to be an open blockchain platform to support the development and use of decentralised applications. Unlike Bitcoin, the programming language available on the Ethereum platform is Turing complete, so that general applications can be run on what the founders call a 'world computer'.

At the core of the Ethereum concept are two types of accounts:

- ***Externally Owned Accounts (EOAs)*** which are controlled by private keys. A private key is a cryptographic mechanism allowing for individuals to sign a transaction which has been secured by a corresponding public key. EOAs are controlled by individual users or organisations.
- ***Contract Accounts,*** also termed 'Smart Contracts' are controlled by contract code and are activated by EOAs.

[5] https://motherboard.vice.com/en_us/article/aek3za/bitcoin-could-consume-as-much-electricity-as-denmark-by-2020

When *ether*, the currency used within Ethereum, is sent from an EOA to a Contract Account, the contained program is executed. This can result in further transactions and payments and additional Smart Contracts being invoked. Through these chains of invocation, connected Smart Contracts form the basis of Ethereum applications called 'dApps' (distributed applications). A number of high-level languages exist for Smart Contracts including Solidity (similar to C/JavaScript), Serpent (similar to Python) and LLL (a low-level Lisp-like language).

One Smart Contract that has attracted attention recently is ERC721 ("Ethereum Request for Comments" #721).[6] This Smart Contract was developed after the Crypto Kitties[7] game became very popular. Players of Crypto Kitties can collect, trade and breed cats, with all activity running on the Ethereum blockchain. At its height, the platform raised $12M of venture capital and the most popular cats were valued at over $100,000. ERC721 extends ERC 20, a Smart Contract for representing tradeable tokens (such as coins). The key contribution of ERC721 is to represent *non-fungible* tokens - tokens which are each unique and therefore not interchangeable. Below we outline how we have extended this Smart Contract in our own work.

The Interplanetary File System (IPFS)[8] is a peer-to-peer distributed file system which is in some ways analogous to the Web, but which uses BitTorrent techniques for exchanging data. Each stored file is indexed by its hash with the indexes stored on network nodes. Human readable decentralised file naming is supported through IPNS analogous to DNS. The benefits brought by IPFS include content-based addressing, increased speed of data delivery over networks since large files can be transported in parallel, easy replication of valuable data, and network resilience in areas of low connectivity. IPFS has often been used together with blockchains as a data storage area and one recent joint venture is Filecoin[9] whereby peer data hosters are paid.

4.2 Verified Claims and Self-sovereign Identity

The Verifiable Claims Working Group (VCWG) aims to make the verifiable exchange of claims easier and more secure on the Web – in particular, how to make claims cryptographically secure, privacy respecting, and automatically verifiable. The main roles in the use of claims and their relationship are outlined in Figure 2, taken from (Sporny and Longley 2017)):

- ***Holder*** - the owner or controller of a number of verifiable claims, for example a student with qualifications or a citizen with personal attributes (e.g. age) or rights (e.g. to work).
- ***Issuer*** - creates verifiable claims, each connected to a specific subject, and then transmitted to claim holders. Examples here include universities (for qualifications) and governments.
- ***Inspector*** verifier - processes all received verifiable claims. Examples include employers and staff managing national borders.

Identifier registry - mediates the creation and verification of entities about which claims are made. Examples may include university records, organisational employee databases and governmental citizenship data stores. It is acknowledged that distributed ledgers can play the role of identifier registries in decentralised contexts. Standard identity management systems are based on centralised authorities such as certificate authorities or domain name registries. These are in some cases perceived as contributing to some of the problems outlined above, that is, contributing to the over centralisation of data. A partial community response to this has centred around the notion of 'self-sovereign identity' whereby users own, control, and manage their data. The main mechanism for this is through the following two main steps. Firstly, using a special app - an identity wallet - the user generates a public/private key pair which is unique and stored locally. Secondly, users request attestations using their public key from authorities and these are digitally

[6] http://erc721.org/

[7] https://www.cryptokitties.co/

[8] https://ipfs.io/

[9] https://filecoin.io/

signed by and stored on the wallet as well. An identity check from an identity authority may be conducted using mobile phone and passport to verify a user's identity. A blockchain can serve as a public trusted store for public keys. The main benefits associated with self-sovereign identity are that users maintain ownership and control over their data. Users can share attestations which can be shaped to minimise the exposure of private data. For example, a citizen may have an attestation, from an appropriate governmental agency, that they are 'over the legal drinking age' independent of any stored record of birth date.

Building on distributed ledger technologies, such as blockchains, the W3C Decentralised Identifiers Community Group have recently generated a data model and syntax for Decentralised Identifiers (DIDs) (Reed et al. 2018). DIDs are a new type of identifier for verifiable self-sovereign digital identity under the control of a DID subject, independent from any centralized authority. Each DID Document is comprised of three parts. *Cryptographic material* enables documents to be authenticated through a suite of *authentication suites. Service endpoints* support the provision of decentralised identity management services using the data contained in the DID.

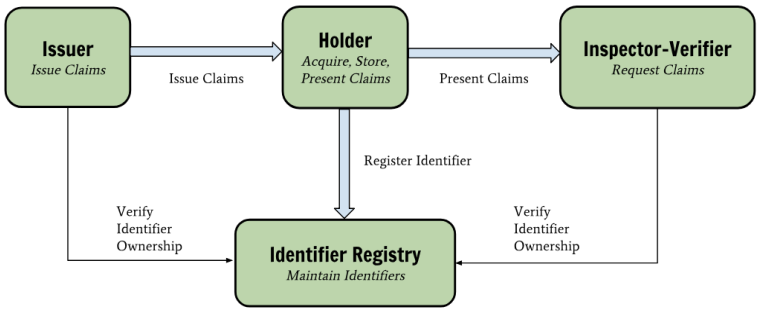


Figure 2: An example type of scenario for the use of verifiable claims. Taken from (Sporny and Longley 2017)

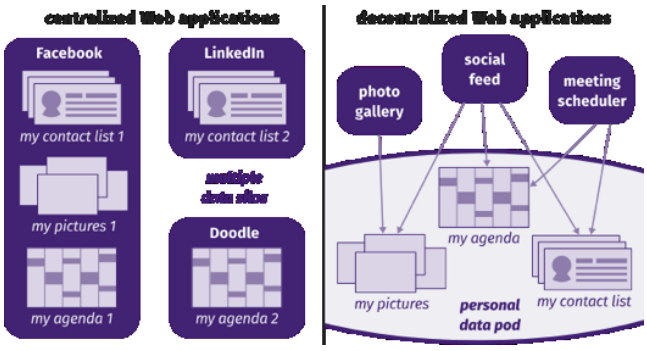


Figure 3: An overview of Solid concepts (taken from (Verborgh 2017))

4.3 Decentralised Linked Data Approaches

The Linked Data model supports, by design, the capacity for distributed or decentralised data systems. A single SPARQL query may, through the SERVICE keyword, pull data from multiple independent sources and aggregate it to produce a single set of results. Linked Data approaches in general, however, are agnostic with regard to distribution and decentralisation and can be used to implement many variations along these spectra. In terms of approaches specifically designed with decentralization in mind, we discuss Solid and Linked Data Fragments.

Solid (Lee 2018b) is a new initiative from the inventor of the Web which aims to enable the decentralisation of personal data. As can be seen in Figure 3, each Solid user has their own personal data which is called a data pod. Users are free to choose who hosts the data - whether to host personally or with

a third party of choice. In the Solid scenario, applications are decoupled from the data they consume, interacting through APIs, with user permission, avoiding vendor lock-in and allowing users to easily switch between applications and servers without loss of service. An additional benefit is that newcomers can innovate and develop new applications for users to try out. As stressed in an open letter (Lee 2018c), Solid is driven by the principle of "personal empowerment through data".

According to Verborgh et al., (Verborgh et al. 2016) a Linked Data Fragment is a subset of an RDF based knowledge graph computed by some means. As stated by the authors the problem with most approaches to obtaining Linked Data Fragments outside of a centralized context is that the burden of computation rests entirely with the client, for example, if the data is downloaded, or with the server, for example, using SPARQL, as shown at the top of Figure 4. As shown at the bottom of Figure 4, Triple Pattern Fragments (Verborgh et al. 2016) allow the computation to be shared between a server and client. In essence the approach lowers the burden on RDF servers since the only computation required is on the simple matching of triples. The Triple Fragment Server returns matching triples in reasonably sized fragments (100 triples approximately) and, in addition, metadata and controls are returned. The metadata includes an estimate of the number of triples that match the given triple pattern. Controls enable clients to retrieve further triples from the same knowledge graph. A Linked Data Fragments client supports standard SPARQL queries over any returned triples. Triple Pattern Fragments make it easy to query across multiple data stores at once - much as with SPARQL's SERVICE keyword, but with the bulk of the computational work of aggregation moved from server to client. Triple Pattern Fragments offer a specific advantage for decentralised querying compared to SPARQL, in that it is no longer *required* to know in advance which pieces of data are to be found on which federated endpoint.

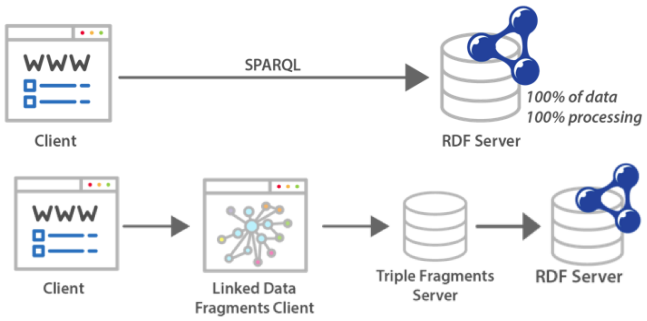


Figure 4: At the top, a standard SPARQL scenario where an RDF server is responsible for hosting and processing all the data. At the bottom, a combination of a Linked Data Fragments client and a Triple Fragments server allows the burden of computation to be shared between a client and server.

4.4 LinkChains

In designing LinkChains, our overall goal has been to bring together the benefits accrued from LD approaches together with decentralised approaches, as embodied by distributed ledgers and IPFS. As can be seen in Figure 5, the LinkChain architecture has three main parts: a user-controlled private storage area; a decentralised public and private data storage area; and a blockchain.

We currently use Solid personal data pods to store private data solely under user control (complete control over the pod-stored data), although LinkChains will talk to any Linked Data Platform.[10] Read/write access to Solid pods is granted through the Solid auth client (Lee 2018b). Non-private or encrypted data is stored within a decentralised data storage system, as opposed to directly on a blockchain, as most public blockchains require payment for data storage and do not store large volumes of data efficiently. We use a blockchain instead to record and verify claimed data attributes, and to facilitate the transfer of data as an asset.

[10] https://www.w3.org/TR/ldp/

A LinkChain Smart Contract enables data on the blockchain to be written or queried. The key functionality provided through our Smart Contract is the ability to store LinkChain Trust Tokens (LCTT) - tokens verifying a set of attributes of a Linked Data set. In particular, a LCTT provides signed verification that declared attributes hold for a Linked Data set. LCTT are created by extending the ERC 721 token standard described above. The main parts of this LinkChain architecture are controlled through a Web/mobile application or a client using the APIs and interfaces provided by the LinkChain platform.

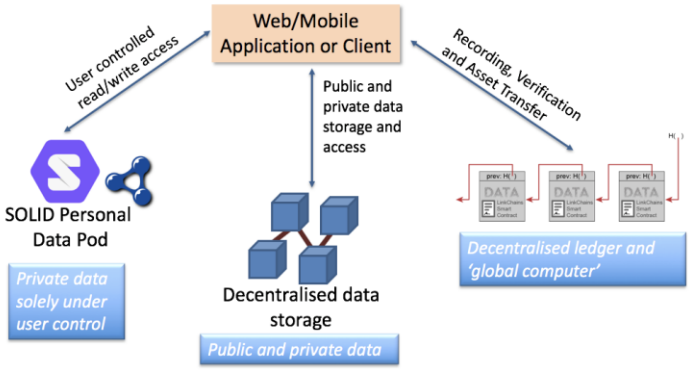


Figure 5: The overall LinkChain architecture which comprises of three main parts: storage of private data under user control; decentralised public and encrypted private data; and a blockchain for recording, verification and asset transfer.

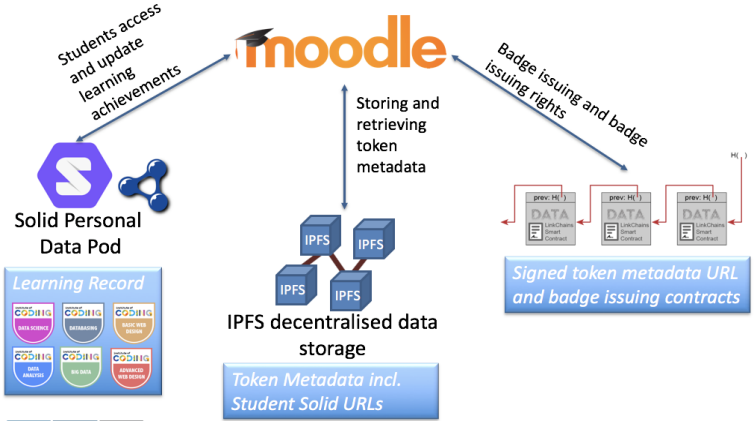


Figure 6: An instantiated LinkChain architecture supporting the storage of personal learning records including Open Badges.

Figure 6 shows how we instantiated the above generic LinkChain architecture to store and verify lifelong learning records. This ongoing work is a part the Institute of Coding[11], an initiative from the UK government to tackle the UK digital skills crisis.[12] When a student completes a particular assignment using the Open University's Virtual Learning Environment (VLE), an Open Badge (based on the IMS Open Badge standard)[13] is automatically issued which students can consequently claim and have stored within their Solid data pod. Additionally, a digitally signed hash of the badge is encoded within an LCTT placed

[11] https://instituteofcoding.org/

[11] https://londonlovesbusiness.com/institute-of-coding-launches-to-tackle-uk-digital-skills-crisis/

[13] https://www.imsglobal.org/cc/statuschart/openbadges

onto the LinkChain blockchain and a token metadata file is stored on IPFS. Placing the metadata on IPFS rather than onto the blockchain reduces cost and increases the overall data handling efficiency. Figure 7 shows a screen snapshot of the student interface used when the system was deployed in our 2018 Summer of Code online school.[14] We can see in Figure 7 that the student has selected her Super Badge, given for successfully completing all nine assignments, and can view the details of the blockchain badge representation. Figure 7 also shows that each element of the Open Badge, including *assertion*, *recipient* and *evidence*, has its own distinct blockchain representation encoding data such as the date of issue and whether the badge has been revoked or has expired.

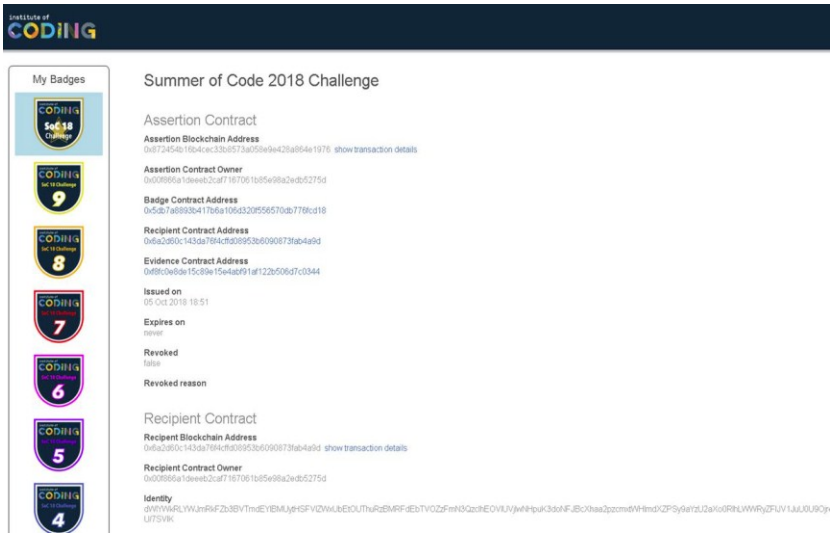


Figure 7: A portion of the interface for OU students viewing their blockchain badges on the 2018 Summer of Code course.

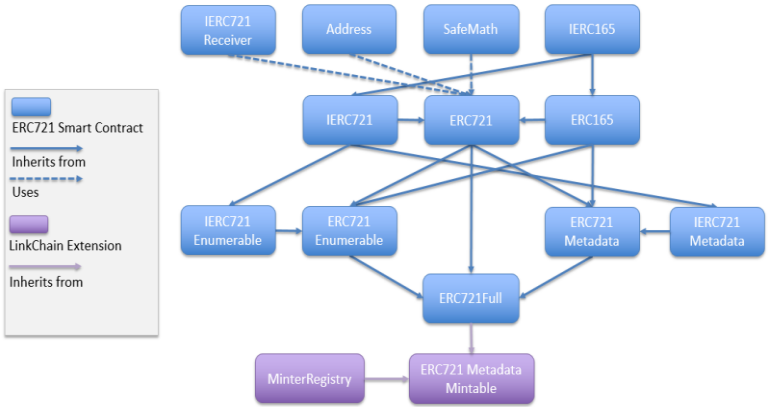


Figure 8: Our extensions to the ERC 721 Smart Contract to support our LinkChains implementation

Figure 8 shows ERC721, including the extension we created for issuing the non-fungible tokens within our LinkChain platform. Each rectangle in the image represents a distinct Smart Contract and the set of blue rectangles together form the ERC721 Smart Contract Template labelled ERC721Full. This template is part of OpenZeppelin[15], a framework of reusable smart contracts for Ethereum and Ethereum Virtual Machine

[13] http://www.open.ac.uk/about/teaching-and-learning/esteem/projects/themes/supporting-students/summer-code

[14] https://github.com/OpenZeppelin/openzeppelin-solidity/blob/master/contracts/token/ERC721/ERC721.sol

(EVM) based blockchains (i.e. blockchains which use the EVM as an execution platform). ERC-721 defines a minimum interface a smart contract must implement to allow unique tokens to be managed, owned, and traded. IERC721 is an interface which defines the functions used in the smart contracts and is used as a way to implement those functions. The "address" smart contract is used to separate account addresses and smart contract addresses. The SafeMath smart contract makes sure that the mathematical operations performed in a smart contract are safety checked which can revert on error. ERC165 standard interface detection is used to expose the interfaces that an ERC721 smart contract supports. The ERC721 Enumerable extension provides functionalities to sort through the tokens easily, whereas the Metadata extension enables a token to have a contract name, symbol and some extra data that makes it unique.
The two smart contracts MinterRegistry and ERC721MetadataMintable (highlighted in purple) together form the core of our LCCT token implementation. MinterRegistry contains a list of addresses which are permitted to mint, i.e. are allowed to issue new ERC721 tokens. ERC721MetadataMintable issues LCCT tokens upon receiving legitimate minter requests in the form of metadata representing the address of the minter, the address that the tokens created are to be sent to, the address of the token contract and the token ID. On ERC721MetadataMintable being invoked, the smart contract authenticates the minting address and issues the LCCT token to the specified address.

5 Evaluation according to FAIR TRADE

We evaluate these principles by considering a number of decentralised data systems and identifying which of the principles, and subprinciples, apply in each case, and to what degree. The goal is to test their *expressive power*: in applying these principles, can we describe significant differences with regard to various systems and approaches? The choice of systems to evaluate is diverse, including systems with significantly different goals and designs, deliberately to illustrate this expressive power.

5.1 Solid

Solid accrues many of its positive scores with respect to the FAIR TRADE framework due to the fact it is a set of conventions and tools based on Linked Data principles and more generally as far as possible on existing W3C standards and protocols. Specifically, the use of URIs supports *findability* through the provision of a unique and persistent identifier. *Accessibility* is supported through the return of RDF based metadata, understandable to both humans and machines, available through HTTP requests and SPARQL queries. Metadata in a formal knowledge representation language such as RDF also supports *interoperability*. The PROV-O provenance vocabulary[16] and the Linked Data Rights vocabulary[17] can support *re-usability*.
We take the relevant network, in the case of Solid, to be the collection of all data pods. *Distribution* and *decentralisation* of data are therefore satisfied for all subprinciples. Each is a distinct per-user Web server allowing users in principle to be completely *autonomous* in terms of data control (AU1), but without a stronger model for identity, AU2 does not hold. Solid does not satisfy any aspects of the *trusted* dimension. This highlights the fact that there is no inbuilt mechanism for verifying that any claims made within a data pod are true. The Verifiable Claims data model can be used to express claims, but additional infrastructure would be required to support verification.

5.2 DBpedia Infrasructure

DBpedia (Lehmann et al. 2015) is, at its core, a machine-readable formal representation of the various Wikimedia projects including Wikipedia. Data within DBpedia is served as Linked Data, enabling semantic query processing. The English version of the DBpedia knowledge base contains over 4.5 million objects, over 4.2 million of which are classified within the DBpedia ontology. DBpedia is not, in itself, of course, a

[16] https://www.w3.org/TR/prov-o/

[17] http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/

data platform; it is a dataset. Nonetheless, there is an infrastructure surrounding its creation, management, and publication, including a central DBpedia site[18] and a network of mirrors. We discuss this infrastructure here.

As it is founded on Linked Data, DBpedia is positively assessed against all the FAIR criteria. Every DBpedia data item has a unique dereferencable URI which contains data and metadata represented in RDF satisfying the *findability*, *accessibility* and *interoperability* dimensions. All data in DBpedia is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License supporting the *re-usability* principle.

Trust is generated from the fact that the data has been extracted from Wikipedia, and that Wikipedia generates trust for the information it holds from its inbuilt consensual and social mechanisms.[19] However, the "automatic" aspects of T1 and T2 are not satisfied. DBpedia data is centralised, and ultimately controlled by the DBpedia Board of Trustees,[20] which currently has eight members; data is not *autonomous* with respect to contributors or users in the senses of AU1 and AU2. DBpedia provides tools to make it easy to maintain a live mirror,[21] however, synchronisation is one-way (central data is pushed to the mirrors); D1 and E2 are satisfied, but D2 and E1 are not.

5.3 Linked Data Fragments

As with the systems described above, Linked Data Fragments passes all the FAIR criteria as it is based on Linked Data, but only partially satisfies the new criteria. As with Solid, there are no inbuilt mechanisms to facilitate *trust* of any provenance for (T1), or claims made about (T2), returned data. Following the core aim of Linked Data Fragments, the platform facilitates the distribution of data between clients and servers - primarily to enable the sharing of computation association with processing. Linked Data Fragments certainly supports both D1 and D2 with regard to *distribution*, and E1 and E2 with regard to *decentralisation*. With regard to *autonomy*, data on a Linked Data Fragments server is controlled by whomever controls the server. and there is no standard means of controlling data visibility. AU1 does not, therefore, hold, nor does AU2: there is no consideration of user identity in the standard at all.

5.4 Bitcoin

Bitcoin core can be used to find transaction/exchange data. But, the metadata or message embedded in the OP_RETURN part of the bitcoin transaction does not lend itself to findable data. Bitcoin as a platform contains no internal indexing or search mechanisms - all discovery is down to external tools which are fixed, or domain specific, or proprietary (e.g. Blockchain.com)[22]; thus transaction/exchange data is *findable*, but data embedded within a Bitcoin transaction is not. As Bitcoin is a public blockchain, all data is *accessible* from the Bitcoin blockchain at any time. No format is prescribed for representing embedded data on Bitcoin, nor is there support for vocabularies; embedded data in a Bitcoin transaction is in general is not *interoperable*. The fact that every Bitcoin transaction including the embedding of data is signed by a private key means that data has at least some verified provenance. This makes embedded data in a Bitcoin transaction *re-usable* to some extent, although it has no standard license, or licensing infrastructure.

Bitcoin miners are distributed over the world and every node has a copy of the entire blockchain. The combination of peer data replication, the proof of work consensus mechanism, and immutability means that Bitcoin data can be *trusted* – neither provenance metadata (T1) nor data contents (T2) can be modified once published. *Autonomous* control of data directly on-chain is limited; data cannot be deleted from the

[18] http://dbpedia.org

[19] https://en.wikipedia.org/wiki/Wikipedia:Consensus

[20] https://wiki.dbpedia.org/board

[21] https://github.com/dbpedia/dbpedia-live-mirror

[22] https://www.blockchain.com/explorer

blockchain, although some control over visibility can be implemented using encryption. Hashes with pointers to off-chain data offer greater potential for autonomy, as does the use of multiple anonymous accounts via a management wallet, but these, particularly the former, require infrastructure going beyond Bitcoin itself. We cannot therefore say that it satisfies either of the autonomy subprinciples The Bitcoin ledger is globally *distributed*: every node has a copy of the full blockchain (D1) and every node executes every script (D2), and fully *decentralised* (E1 and E2). There have been concerns within the Bitcoin community that an individual or consortium of miners could take over the network by gaining more than 50% of the total hashing power and thereby the ability to rewrite history. Thus far, this has not happened; at the time of writing, all mining groups have less than 20% control of the overall network.[23]

5.5 Ethereum

As Ethereum and Bitcoin share similar features, its evaluation results according to FAIR TRADE are similar to Bitcoin. A noteworthy difference is with regard to *autonomy*: the availability of smart contracts gives Ethereum increased power in terms of what the blockchain network itself can support; the autonomy subprinciples themselves may potentially be easier to support on Ethereum with a self-sovereign identity and Verifiable Claims infrastructure using smart contracts. Nonetheless, further work, and off-chain infrastructure, would be required in order to support both AU1 and AU2 fully.

5.6 LinkChains

LinkChains is designed to combine the user-focused Linked Data design of Solid with a blockchain-backed trust layer, and thus inherits traits from both Solid and Ethereum. That it meets more of the FAIR TRADE criteria than other systems here is not a surprise; a FAIR TRADE analysis played a role in the development process. Rather, it provides an example of how this framework can focus efforts by making explicit the principles for high-quality decentralised data management. The key element *not* yet present in the LinkChains design for full FAIR TRADE compliance is support for *autonomous* identity, although we are currently exploring solutions for self-sovereign identity approaches to address this gap.

Table 1: Evaluation according to FAIR TRADE

	Solid	DBpedia	LinkChains	Bitcoin	Ethereum	Linked Data Fragments
Findable	Via LD	Yes	Via Solid	No internal index or search	Similar to Bitcoin.	Yes
Accessible	Via LD	Yes	Via Solid	Yes	Yes	Yes
Interoperable	Via LD	Yes	Via Solid	Data in arbitrary format	Data in arbitrary format	Yes
Re-usable	Via LD	Yes	Via Solid	No licensing.	No licensing.	Yes

[23] https://www.blockchain.com/en/pools

Trusted	Neither T1 nor T2.	Neither T1 nor T2; not automatic.	T1 and T2, via blockchain.	T1 and T2, by blockchain design.	T1 and T2, by blockchain design.	Neither T1 nor T2.
Autonomous	AU1, not AU2.	Neither AU1 nor AU2.	AU1 via Solid, not AU2.	Neither AU1 nor AU2.	Neither AU1 nor AU2.	Neither AU1 nor AU2.
Distributed	D1 and D2.	D1, not D2	D1 and D2.	D1 and D2	D1 and D2.	D1 and D2
Decentralised	E1 and E2.	Not E1. E2.	E1 and E2	E1 and E2.	E1 and E2.	E1 and E2.

6 Conclusions

As we stated at the start of this paper, a growing number of societal concerns have been raised over the last few years related to how the centralisation of data has led to the loss of user control and the dangers associated with data misuse. A number of different technical communities associated with peer-to-peer and Web technologies have been responding to this in part with a variety of new approaches and platforms to decentralising data. The contribution of this paper is to provide a framework to begin to compare systems, based on an explicit set of principles relating to decentralised data management.

Above, we have assessed a range of data solutions for adherence to the FAIR TRADE principles. As well as a measure of quality in data stewardship, the exercise of assessment showed that these principles, even when not fully followed, provide a good characterisation of the properties of the data or platform being assessed. One can see at a glance, for example, that the DBpedia project is concerned with providing a common resource -- concerns relating to autonomy do not apply when the goal is to elicit effectively *donations* of information for public use, but those relating to trust, openness and reuse do apply, with distribution only as required to ensure the accessibility of data. The Bitcoin and Ethereum platforms meet most of the TRADE standards, except that of being Autonomous, but not FAIR, reflecting their development from cryptocurrencies and, in Ethereum's case, distributed computation - their use as decentralised data platforms coming later. Solid and Linked Data Fragments are evaluated to be very similar, reflecting their shared concerns and foundations on Web data standards. It is perhaps unsurprising to see that the platforms which most fit the FAIR TRADE principles are those stemming from a field in which the focus is on data publication and standards. LinkChains, although a work in progress, is being designed to be FAIR TRADE from the outset, and is based on a combination of other technologies, each of which follows some of the principles. In particular, combining Solid's focus on individual data with the possibilities for trusted and verifiable data based on blockchains, distributed identity and Verifiable Claims has the potential to provide a flexible platform with best-practice principles in data management in the decentralised sphere. Note that these systems did not stretch the descriptive capacity of the FAIR TRADE principles: as noted earlier, each of the presented principles can cover a range of possible useful or interesting behaviours which a distributed data solution could display. As novel technologies arise, we expect this framework to continue to provide a useful common vocabulary for assessing and comparing decentralised data systems. By taking a forward-looking approach to data management standards in a decentralised context, we aim to contribute to the establishment of best practice and well-motivated technical approaches early in the development of the field, at a time when their adoption is significantly easier than in a more entrenched technical landscape.

REFERENCES

Asano, Y., Hidaka, S., Hu, Z., Ishihara, Y., Kato, H., Ko, H.-S., Nakano, K., Onizuka, M., Sasaki, Y., and Shimizu, T., 2018. A View-based Programmable Architecture for Controlling and Integrating Decentralized Data. *arXiv preprint arXiv:1803.06674*.

Bureau, C. F. P., 2015. *Data Point: Credit Invisibles* [online]. Available from: https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

Buterin, V., 2013. *Ethereum White Paper* [online]. Available from: https://github.com/ethereum/wiki/wiki/White-Paper [Accessed 10 Dec 2018].

Curran, D., 2018. Are you ready? Here is all the data Facebook and Google have on you. *The Guardian* [online], 2018. Available from: https://www.theguardian.com/commentisfree/2018/mar/28/all-the-data-facebook-google-has-on-you-privacy [Accessed 10 Dec 2018].

Fact, F., 2018. Windrush generation: what's the situation? [online]. Available from: https://fullfact.org/immigration/windrush-generation/ [Accessed 10 Dec 2018].

Grierson, J., 2018. Windrush row: 63 people could have been wrongly removed. *The Guardian* [online], 2018. Available from: https://www.theguardian.com/uk-news/2018/may/15/windrush-row-63-people-could-have-been-wrongly-removed-says-javid [Accessed 10 Dec 2018].

Guardian, T., 2015. The Cambridge Analytica Files. *The Guardian* [online], 2015. Available from: https://www.theguardian.com/news/series/cambridge-analytica-files [Accessed 10 Dec 2018].

Lee, T. B., 2018a. Interview in Vanity Fair. [online]. Available from: https://www.vanityfair.com/news/2018/07/the-man-who-created-the-world-wide-web-has-some-regrets [Accessed 10 Dec 2018].

Lee, T. B., 2018b. SOLID Project Website. [online]. Available from: https://solid.mit.edu/ [Accessed 10 Dec 2018].

Lee, T. B., 2018c. One Small Step for the Web. [online]. Available from: https://www.inrupt.com/blog/one-small-step-for-the-web.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., and Auer, S., 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2), 167–195.

Merkle, R. C., 1980. Protocols for public key cryptosystems. *In*: *Security and Privacy, 1980 IEEE Symposium on*. IEEE, 122.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., and Wilkinson, M. D., 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37 (1), 49–56.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system.

Özsu, M. T. and Valduriez, P., 2011. *Principles of distributed database systems*. Springer Science & Business Media.

Reed, D., Sprony, M., Longley, D., Allen, C., Grant, R., and Sabadello, M., 2018. Decentralized Identifiers (DIDs) v0.11 Data Model and Syntaxes for Decentralized Identifiers (DIDs). *W3C* [online]. Available from: https://w3c-ccg.github.io/did-spec/ [Accessed 10 Dec 2018].

Schneider, F. B. and Trustworthiness, C. on I. S., 1999. *Trust in cyberspace*. National Academy Press Washington, DC.

Sporny, M. and Longley, D., 2017. *Verifiable Claims Data Model and Representations* [online]. Available from: https://www.w3.org/TR/2017/WD-verifiable-claims-data-model-20170803.

Third, A. and Domingue, J., 2017. LinkChains: Exploring the space of decentralised trustworthy Linked Data.

Verborgh, R., 2017. Paradigm shifts for the decentralized Web. [online]. Available from: https://ruben.verborgh.org/blog/2017/12/20/paradigm-shifts-for-the-decentralized-web/ [Accessed 10 Dec 2018].

Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., and Colpaert, P., 2016. Triple Pattern Fragments: a low-cost knowledge graph interface for the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37, 184–206.

Walport, M., 2015. *Distributed Ledger Technology: beyond block chain* [online]. Government Office for Science. Available from: https://pubs.acs.org/doi/pdf/10.1021/acsaem.8b00240%0Ahttp://pubs.acs.org/doi/10.1021/acsaem.8b00240.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., and Bourne, P. E., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Wilkinson, M. D., Verborgh, R., da Silva Santos, L. O. B., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., van Mulligen, E. M., and Ciccarese, P., 2017. Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, 3, e110.

Enabling Automatic Discovery and Querying of Web APIs at Web Scale using Linked Data Standards

Franck Michel
Université Côte d'Azur, CNRS, Inria, I3S, France
franck.michel@cnrs.fr

Catherine Faron-Zucker
Université Côte d'Azur, Inria, CNRS, I3S, France
faron@i3s.unice.fr

Olivier Corby
Université Côte d'Azur, Inria, CNRS, I3S, France
olivier.corby@inria.fr

Fabien Gandon
Université Côte d'Azur, Inria, CNRS, I3S, France
fabien.gandon@inria.fr

ABSTRACT

To help in making sense of the ever-increasing number of data sources available on the Web, in this article we tackle the problem of enabling automatic discovery and querying of data sources at Web scale. To pursue this goal, we suggest to (1) provision rich descriptions of data sources and query services thereof, (2) leverage the power of Web search engines to discover data sources, and (3) rely on simple, well-adopted standards that come with extensive tooling. We apply these principles to the concrete case of SPARQL micro-services that aim at querying Web APIs using SPARQL. The proposed solution leverages SPARQL Service Description, SHACL, DCAT, VoID, Schema.org and Hydra to express a rich functional description that allows a software agent to decide whether a micro-service can help in carrying out a certain task. This description can be dynamically transformed into a Web page embedding rich markup data. This Web page is both a human-friendly documentation and a machine-readable description that makes it possible for humans and machines alike to discover and invoke SPARQL micro-services at Web scale, as if they were just another data source. We report on a prototype implementation that is available on-line for test purposes, and that can be effectively discovered using Google's Dataset Search engine.

KEYWORDS

SPARQL, Web API, discovery, dataset, Web service, Linked Data

ACM Reference Format:
Franck Michel, Catherine Faron-Zucker, Olivier Corby, and Fabien Gandon. 2019. Enabling Automatic Discovery and Querying of Web APIs at Web Scale using Linked Data Standards. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317073

1 INTRODUCTION

As data sources multiply all over the Web, it becomes increasingly crucial to enable the automatic discovery of sources suitable to answer a certain query. Multiple portals and services allow to search datasets based on metadata, such as the manifold CKAN-based portals[1] like Datahub[2] and the data portals of European states and institutions. Some portals specialize in specific dataset formats or interface technologies. For instance, ProgrammableWeb.com[3] registers Web APIs, a loosely defined category of lightweight Web services also referred to as REST-like or Lo-REST [25] services, while LODAtlas [26] and SPARQLES [31] focus on RDF datasets and SPARQL endpoints respectively.

Even though some of these portals have gained significant popularity due to the large number of datasets that they index, they suffer relentless flaws. Firstly, they are centralized registries with a somehow restricted scope. Consequently, potential data consumers may have to query several portals one by one, accommodating the various query interfaces, to discover suitable datasets. Secondly, in many cases, datasets are manually registered and annotated by dataset producers, thereby raising concerns about outdated metadata or deprecated services. Thirdly, metadata-based search results have a limited relevance. Typically, searching datasets by keywords and data formats is a first step in the discovery process, but a potential consumer needs deeper insight in the data themselves and the technical interfaces available to query the dataset. In this respect, WSDL-based semantic Web services (e.g. OWL-S [4] or SAWSDL [8]) tackled this question with a thorough description of the exchanged messages, yet often failing to describe the actual dataset being queried. Besides, they were better suited to the controlled environment of companies [20] than the open environment of the Web. By contrast, the VoID vocabulary [1] can help describe RDF datasets with regards to vocabularies, classes and properties used, links to other datasets, etc. But it does not address the description of what properties a resource may typically have nor how the resources relate to each other, which are key criteria in the discovery and selection of datasets.

To spur and enable automatic discovery and consumption of datasets at Web-scale, we believe that a few principles should drive future research and developments.

(1) **Metadata-based search is not enough**. As we pointed out above, metadata-based search using e.g. keywords, data formats, vocabularies or even classes and properties used in an RDF dataset, is just a first step in the discovery process. For example, assume a biologist wants to develop a software agent capable of browsing Linked Data and gathering photos

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317073

[1] https://ckan.org/about/instances/
[2] https://datahub.ckan.io/
[3] https://www.programmableweb.com/

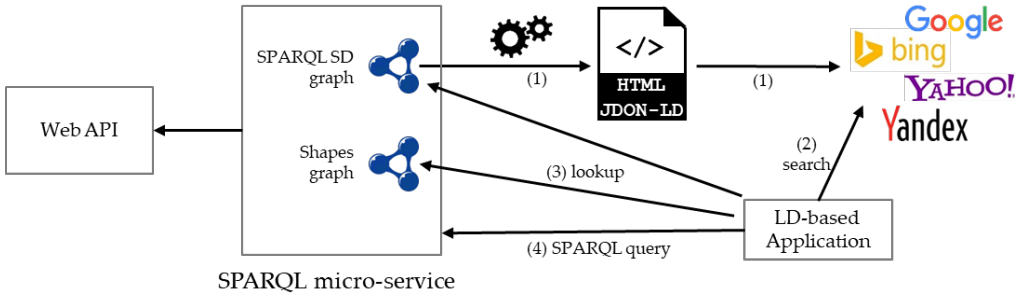


Figure 1: Envisaged use case spanning the indexing, discovery and invocation of a SPARQL micro-service.

related to biological species. The agent may submit several queries to repositories such as LODAtlas looking for datasets whose textual description contains keywords "photo" and "biodiversity", or those whose VoID description (if any) mentions classes representing photographies and biological taxa. Within the matching datasets, however, nothing guarantees that photographic resources do actually depict biological species; photographies may well be scans of academic papers related to the species. Hence, the agent has no choice but to query the dataset in order to get insight into it and find out if it matches the search. This simple example illustrates the lack of in-depth semantic description of datasets, that would consist of the resources (what are the actual properties of photographic and taxonomic resources) along with their mutual relationships.

(2) **The discovery of datasets at Web-scale should leverage the power of Web search engines**. Major search engines such as Google, Yahoo and Bing crawl and index an unprecedented breadth of information every day. They already harvest the content of specialized open data portals, in particular by taking advantage of the growing use of the Schema.org vocabulary [12]. Google has recently opened a beta service specifically dedicated to dataset search[4]. Therefore, despite concerns raised by the Web centralization effect of search engines, it is worth studying how we can take advantage of their services to enable the discovery and querying of datasets at Web scale.

(3) **The description of datasets and their query services should rely on well adopted (de-facto) standards**. Enabling the automatic discovery and querying of datasets at Web-scale means that, at some point, a consensus should be reached with respect to technologies and practices. Such a consensus may emerge only if the selected approaches put little constraints on and require little efforts from those in charge of describing datasets, publishing and maintaining query services thereof. This means relying on existing, well-adopted standards or de-facto standards.
In terms of semantic description, existing vocabularies should be leveraged, ranging from mature and widely used W3C standards to de-facto standards such as Schema.org that benefits from a large and growing adoption even though it still lacks terms in many domains. Additionally, selected approaches should enjoy sufficient and appropriate tooling with APIs in various programming languages. Such tools should be relatively simple in the sense that (i) they should not require a long learning curve from developers, and (ii) they should be easy to deploy and maintain. In this respect, the example of WSDL-based semantic Web service frameworks is inspiring: their deployment and operation required significant efforts that only companies with solid IT services were ready to invest [25]. But when seeking Web-scale adoption, such perceived complexity would have a counterproductive effect.

[4]https://www.blog.google/products/search/making-it-easier-discover-datasets/

In a previous work, we defined the SPARQL Micro-Service architecture [23] aimed at querying Web APIs using SPARQL [14], thus bridging the Linked Data and Web API worlds. We suggested that this approach could foster the emergence of an ecosystem of SPARQL services published by independent providers, allowing Linked Data-based applications to glean pieces of data from a wealth of distributed data sources, in a scalable and reliable manner.

In this article, we present further exploratory works aimed at applying the principles set out above, and thereby make SPARQL micro-services effectively discoverable and queryable at Web-scale. We describe and explain our architectural and modeling choices. Let us however underline that alternative choices may be figured out, driven by different incentives or trade-offs. We touch upon these considerations in the last section.

Envisaged use case. Figure 1 outlines the main steps of a typical use case as we see it, along with the main choices that we made. A SPARQL micro-service produces a Web page (step 1) whose primary goal, beyond providing developers with appropriate documentation and a testing interface, is to be processed by Web crawlers. It embeds rich markup data, notably based on Schema.org, to enhance indexing and help search engines yield more accurate results. The Web page is generated dynamically from the service self-description that consists of a SPARQL Service Description (SD) graph [33] and a SHACL shapes graph [16]. Together, they provide various metadata, a description of the graphs that the service typically spawns, the service inputs and outputs and the way they relate to one another. An application willing to carry out a certain task first queries search engines (step 2) for datasets matching certain keywords. From the search results, it extracts and looks up SPARQL endpoint URLs. SPARQL micro-services return an SD document that links to the shapes graph. In turn, the application fetches the shapes graph that allows verifying whether the service is indeed suited for the task

Figure 2: SPARQL micro-service processing workflow.

(step 3). Based on the description of the service inputs, the application can submit an appropriate SPARQL query to the micro-service (step 4).

The rest of this article is organized as follows. Section 2 briefly summarizes the concepts of SPARQL micro-services and presents a quick example. Section 3 then presents the way we describe micro-services in a machine-readable manner. Section 4 focuses on the way micro-services are made discoverable at Web scale. Related works are discussed in section 5 while the last section brings elements of discussion and suggests future leads.

2 BACKGROUND

In [22], we described the SPARQL Micro-Service architectural principles. Later on in [23], we extended this description and reported on several biodiversity-related use cases. In this section, we briefly summarize these previous works.

The SPARQL Micro-Service architecture addresses the problem of combining Linked Data with data from non-RDF Web APIs. A SPARQL micro-service is a lightweight SPARQL endpoint that provides access to a graph generated at run-time. This graph is shaped by the Web API service being wrapped, the arguments passed to the micro-service and the types of RDF triples it is designed to produce. How the arguments are passed to a SPARQL micro-service, and how the Web API response is transformed into a SPARQL result, are implementation choices.

In accordance with the micro-service architecture principles [24], a SPARQL micro-service is typically designed to be loosely coupled (it is deployed independently of other services, possibly using lightweight container technologies such as Docker[5]) and fine-grained: it provides access to a small, resource-centric graph corresponding to a small fragment of the whole dataset served by the Web API.

Interestingly, this architecture can be used to assign dereferenceable URIs to Web API resources that do not have URIs in the first place: a micro-service responds to SPARQL queries by assigning URIs to Web API resources, while other micro-services are designed to dereference these URIs to RDF content. This entails an effective solution to bridge Web APIs, that are designed as closed worlds, with the open world of Linked Open Data.

Implementation. We implemented a lightweight PHP prototype available on GitHub[6] under the Apache 2.0 license. The prototype focuses on JSON-based Web APIs, and expects arguments of a micro-service to be passed as parameters of the service URL's query string. Figure 2 illustrates how a SPARQL micro-service S_μ evaluates a SPARQL query Q. In step 1, S_μ receives query Q and extracts the set Arg_w of arguments from the HTTP query string. In step 2, it invokes the Web API with the arguments in Arg_w, in addition to any other parameter required by the Web API. In step 3, S_μ translates the JSON response into an RDF graph: it carries out a first mapping towards selected vocabularies by applying a JSON-LD profile [30] to the response; the resulting graph G is loaded into a local triple store; if mappings are needed that JSON-LD cannot express, S_μ runs a SPARQL INSERT query that enriches G with additional triples. Finally, S_μ evaluates Q against G and returns the result to the client.

Alternative argument-passing method. In the method described above, the arguments of a SPARQL micro-service are passed as query string parameters rather than RDF terms. One advantage is that it spares creating new terms whenever a Web API-specific argument has no counterpart in existing vocabularies. Nevertheless, a downside is that the semantics of such a SPARQL micro-service differs from that of a standard SPARQL endpoint. Indeed, the SPARQL protocol treats a service URL as a black box, i.e. it does not identify nor interpret URL parameters apart from those specified in the SPARQL protocol itself. By contrast, in a SPARQL micro-service the query string parameters are meaningful arguments that shape the virtual graph being queried. Therefore, since one of our goals in this article is to comply with standards (principle 3), we have recently implemented an alternative method wherein arguments are passed as regular RDF terms of the SPARQL query graph pattern. To illustrate this, we now introduce an example that we shall reuse throughout the rest of this article.

Running example. Let us consider the service of Flickr's Web API that returns a list of photos matching some criteria[7]. We define $S_{\mu f}$ as a SPARQL micro-service[8] that wraps this Flickr service and returns photos of a given biological species. $S_{\mu f}$ takes as argument the species scientific (taxonomic) name, and searches photos matching this name. It abides by the convention that photos of a species should be tagged with the species scientific name formatted as `taxonomy:binomial=<scientific name>`[9]. $S_{\mu f}$ expects the scientific name argument to be passed as the object of the `dwc:scientificName` predicate.

Listing 1 depicts a query, Q_1, that meets this requirement. It aims at retrieving photos depicting species *Delphinus delphis*, the common dolphin. When it evaluates Q_1, $S_{\mu f}$ first extracts the scientific name argument from the graph pattern (highlighted line) and builds the following Web API invocation URL:

```
https://api.flickr.com/services/rest/?
  format=json&method=flickr.photos.search&
  tags=taxonomy:binomial=Delphinus+delphis
```

It then submits this invocation and translates the JSON response into an RDF graph such as the one exemplified in Listing 2. Finally,

[5] https://www.docker.com
[6] https://github.com/frmichel/sparql-micro-service/tree/0.3.1/

[7] https://www.flickr.com/services/api/flickr.photos.search.html
[8] The code of this service is available at https://frama.link/kVhhnE-v.
[9] This is a common convention used on Flickr for biodiversity resources, in particular the Encyclopedia of Life group (https://www.flickr.com/groups/806927@N20) and the Biodiversity Heritage Library (https://www.flickr.com/photos/biodivlibrary).

```
prefix schema: <http://schema.org/>
prefix dwc:    <http://rs.tdwg.org/dwc/terms/>

SELECT ?title ?img WHERE {
 ?taxon a dwc:Taxon;
   dwc:scientificName "Delphinus delphis";
   schema:image [
     a schema:Photograph;
     schema:name ?title;
     schema:contentUrl ?img.
    ].
}
```

Listing 1: Query Q_1 can be submitted to SPARQL micro-service $S_{\mu f}$ in order to retrieve photos of species *Delphinus delphis*.

```
[] a dwc:Taxon;
  dwc:scientificName "Delphinus delphis";
  schema:image <http://example.org/ld/flickr/photo/31173091626>.

<http://example.org/ld/flickr/photo/31173091626>
 a schema:Photograph;
 schema:name "Delphinus delphis 5 (13-7-16 San Diego)";
 schema:contentUrl
  <https://farm6.staticflickr.com/5718/31173091626_88c410c3f2_z.jpg>;
 schema:mainEntityOfPage
  <https://flickr.com/photos/10770266@N04/31173091626>;
 schema:fileFormat "image/jpeg";
 schema:author [
  schema:identifier "10770266@N04";
  schema:url <https://flickr.com/photos/10770266@N04>
 ].
```

Listing 2: Example graph produced by micro-service $S_{\mu f}$ to evaluate query Q_1.

it evaluates Q_1 against this graph, yielding the response exemplified below in the SPARQL Query Results JSON format [28]:

```
{ "head": { "vars": [ "title", "img" ] }, "results": {
    "bindings": [
      { "title": {
          "type": "literal",
          "value": "Delphinus delphis 5 (13-7-16 San Diego)" },
      { "img": {
          "type": "uri",
          "value": "https://farm6.staticflickr.com/5718/ \
                    31173091626_88c410c3f2_z.jpg" }
} ] } }
```

3 MACHINE-READABLE DESCRIPTION OF SPARQL MICRO-SERVICES

Building on the work presented in section 2, we aim at proposing a mechanism that enables a software agent to discover, select and invoke the SPARQL micro-services that are relevant for a certain task. In section 1, we pointed out three principles that, we believe, should help pursue this goal: (1) have rich descriptions of data sources that go beyond common metadata, (2) leverage Web search engines to discover data sources, and (3) rely on well-adopted standards. This section presents the modeling choices we made with respect to principle (1), section 4 deals with principle (2) while principle (3) is transversal to both sections.

A rich SPARQL micro-service description should span two distinct levels further detailed in this section. The **high-level description** consists of metadata about the data being queried (keywords, publisher, license, vocabularies, graphs, etc.), as well as metadata about the micro-service itself (supported operations, result formats, etc.). The **functional description** describes the actions the services carries out: what are the types of resources involved, what are their recommended/expected properties, what are the service arguments and how they relate to the resources.

3.1 High-level Description

To describe SPARQL micro-services, we use SPARQL Service Description (SD) [33] which is both a vocabulary to describe SPARQL endpoints and a method requiring compliant endpoints to return an SD document when their URL is looked up.

Listing 3 depicts a snippet of the SD document (in the Turtle syntax) for the example service $S_{\mu f}$ introduced in section 2. The service is at the same time an instance of the SD Service class and the class of SPARQL micro-services `sms:Service` (line 17). Common metadata are provided lines 19 to 26, such as a name and description, keywords, supported SPARQL language and result formats. A VoID description can also be embedded here, as exemplified in line 29 (the default dataset is stated to be a void:Dataset) and lines 34 to 36[10]. Additional triples are not depicted here for conciseness, such as the service publisher and an example SPARQL query. Note that many more metadata could be provided, such as common dataset profile features [2]. Furthermore, in the implementation we demonstrate here, we wrote the SD document manually. Future works could consider dataset profiling techniques to (at least partially) automate this generation.

The SD document is obtained by looking up the service URL. Content negotiation is supported such that a Web browser will obtain an HTML page, whereas a Linked Data application would typically require one of the supported RDF serialization syntaxes. The SD document itself is a named graph of the dataset served by the SPARQL micro-service (line 31). The interested reader may view the full range of metadata by looking up the named graph URI[11] in a Web browser (this will typically return an RDF/XML representation) or by issuing the following command on a standard Linux system:

```
curl --header "Accept: text/turtle" \
  http://sms.i3s.unice.fr/sparql-ms/flickr/getPhotosByTaxon_sd/
```

3.2 Functional Description

There exist various options to represent the functional description of a service. In section 5 we discuss some of them. As far as SPARQL micro-services are concerned, we choose to leverage several vocabularies for this purpose: SHACL [16], Schema.org and Hydra [19].

SHACL Description of the Dataset. SHACL, the Shapes Constraint Language, is designed for the validation of RDF graphs (called *data graphs*) against a set of conditions expressed in the form of *shapes graphs*. In our context, instead of using a shapes graph G_{sh} *a posteriori* to validate the data graph produced by a SPARQL micro-service, we consider G_{sh} as a specification of the graphs that a SPARQL micro-service can generate.

The shapes graph is linked to the SD document as follows: the default dataset has a default graph that is validated by the shapes

[10] As an alternative, a VoID description could be made available using the *well-known URIs* mechanism, at path `/.well-known/void`.

[11] http://sms.i3s.unice.fr/sparql-ms/flickr/getPhotosByTaxon_sd/ServiceDescription

```
1 prefix xsd:    <http://www.w3.org/2001/XMLSchema#>
2 prefix sd:     <http://www.w3.org/ns/sparql-service-description#>
3 prefix frmt:   <http://www.w3.org/ns/formats/>
4 prefix dct:    <http://purl.org/dc/terms/>
5 prefix shacl:  <http://www.w3.org/ns/shacl#>
6 prefix void:   <http://rdfs.org/ns/void#>
7 prefix hydra:  <http://www.w3.org/ns/hydra/core#>
8 prefix schema: <http://schema.org/>
9 prefix skos:   <http://www.w3.org/2004/02/skos/core#>
10 prefix dwc:    <http://rs.tdwg.org/dwc/terms/>
11 prefix sms:    <http://ns.inria.fr/sparql-micro-service#>
12
13 @base
14   <http://sms.i3s.unice.fr/sparql-ms/flickr/getPhotosByTaxon_sd/>.
15
16 <>
17   a sd:Service, sms:Service;
18   sd:endpoint <>;
19   sd:supportedLanguage sd:SPARQL11Query;
20   sd:feature sd:BasicFederatedQuery, sd:EmptyGraphs;
21   sd:resultFormat
22       frmt:SPARQL_Results_JSON, frmt:Turtle, frmt:JSON-LD;
23
24   schema:name "Search for Flickr photos by taxon scientific name";
25   schema:description "...";
26   schema:keywords "biodiversity", "lifesciences", "photography";
27
28   sd:defaultDataset [
29     a sd:Dataset, void:Dataset;
30     sd:defaultGraph [a sd:Graph; shacl:shapesGraph <ShapesGraph>];
31     sd:namedGraph   [a sd:Graph; sd:name <ServiceDescription>];
32     sd:namedGraph   [a sd:Graph; sd:name <ShapesGraph>];
33
34     void:vocabulary
35         <http://schema.org/>, <http://rs.tdwg.org/dwc/terms/>;
36     void:sparqlEndpoint <>;
37   ];
38
39   dct:source [
40     # Web API service being wrapped by this service
41     a schema:WebAPI; schema:name "Flickr API";
42     schema:url <https://www.flickr.com/services/api/>;
43     schema:potentialAction [
44       a schema:SearchAction;
45       a hydra:IriTemplate;
46       hydra:template "https://api.flickr.com/services/rest/? \
47         format=json&method=flickr.photos.search& \
48         tags=taxonomy:binomial={name}"
49
50       hydra:mapping [
51         hydra:variable "name";
52         schema:description "Taxon scientific name";
53         hydra:required "true"^^xsd:boolean;
54         skos:example "Delphinus delphis";
55
56         # Use either hydra:property or shacl:sourceShape
57         hydra:property dwc:scientificName;
58       ];
59     ];
60   ].
```

Listing 3: Snippet of the Service Description of SPARQL micro-service $S_{\mu f}$.

graph (property `shacl:shapesGraph` lines 30). The shapes graph is itself one of the named graphs of the default dataset (line 32).

A short snippet of the shapes graph corresponding to service $S_{\mu f}$ is given in Listing 4. The interested reader may check the complete shapes graph on GitHub[12] or by dereferencing its URI[13]. It states that an instance of class `dwc:Taxon` (lines 4-5) should have exactly three properties: `rdf:type` with object `dwc:Taxon` (lines 9-10), `shacl:image` whose object should be validated against another shape (lines 12-13) and property `dwc:scientificName` that should have exactly one literal object (lines 17-18). Notice that the graph pattern of query Q_1 (Listing 1) specifically matches these constraints.

Description of the Input Arguments. We now need to characterize the micro-service input arguments, how they are extracted from a SPARQL graph pattern, and how they map to parameters of the Web API wrapped by the micro-service. We define the Web API as the micro-service data source (line 39 of Listing 3). It is typed as a Schema.org `WebAPI` having one potential action of type `SearchAction` (lines 40-44). Note that an alternative is currently being discussed within the Schema.org community, that links `EntryPoint` objects to a `WebAPI` [27]. The search action is also typed as a Hydra `IriTemplate` whose template string is the Web API invocation URL (lines 46-48). Each mapping (lines 50-58) maps a parameter used in the template string to a term of the SPARQL query by pointing to a specific property using `hydra:property`. In our example, the scientific name, denoted "`{name}`" in the template string (line 48), is mapped to property `dwc:scientificName` (line 57). Upon invocation, the service simply reads the value of property `dwc:scientificName` in the graph pattern, and substitutes it with "`{name}`" in the template string.

This solution is simple and concise, but it presents two downsides: (i) `hydra:property` only names a property but does not put any other constraint such as what is the subject of this property, or how many values are allowed; (ii) there is no explicit relationship between the input argument and the shapes graph. Hence, to specify the input arguments more precisely, an alternative is to map the parameter to a property shape of the shapes graph. In our example, this would be expressed by replacing line 57 with the following:

```
shacl:sourceShape <ShapesGraph#NamePropertyShape>;
```

The referenced property shape is defined in Listing 4 (lines 16-18). Not only it instructs that the scientific name should be given by property `dwc:scientificName`, but also that this property should be attached to an instance of the `dwc:Taxon` class and that there should be only one such property.

Advantages of using SHACL. We believe that using SHACL presents two advantages:
(1) SHACL's expressiveness allows denoting complex relationships between resources (e.g. cardinality, predicate paths). Even though this description is schema-based, it is sufficient to enable SPARQL micro-service discovery and selection since, by construction, the shape of generated graphs is know at design time. By contrast, SPARQL federated query engines generally rely on dynamic instance-based statistics because the graphs being queried can hardly be characterized by a static SHACL description. For instance, it would be impossible to define a precise shapes graph of crowd-sourced graphs such as DBpedia.
(2) A SHACL shapes graph is itself an RDF graph. Therefore, a software agent can leverage existing tooling to reason upon it and verify whether the SPARQL micro-service fulfills the agent's goals. As an illustration, we are currently developing a *SPARQL micro-service federated query engine*[14]. Given an input SPARQL query, the engine searches candidate SPARQL micro-services whose inputs are satisfied by the query. It then selects those whose shapes graphs validate some triple patterns of the query, and finally rewrites the input query into a UNION of SERVICE clauses that invoke SPARQL

[12]Complete shapes graph on GitHub: https://frama.link/we_EQWnC
[13]Shapes graph URI: http://sms.i3s.unice.fr/sparql-ms/flickr/getPhotosByTaxon_sd/ShapesGraph

[14]Beta version available at https://frama.link/VWG7r8PF.

```
@base
  <http://sms.i3s.unice.fr/sparql-ms/flickr/getPhotosByTaxon_sd/>.

<ShapesGraph#TaxonShape> a shacl:NodeShape;
  shacl:targetClass dwc:Taxon;
  shacl:property
    <ShapesGraph#NamePropertyShape>,
    [ a shacl:PropertyShape;
      shacl:path rdf:type;
      shacl:hasValue dwc:Taxon ],
    [ a shacl:PropertyShape;
      shacl:path schema:image;
      shacl:node <ShapesGraph#PhotographShape> ];
  shacl:closed true.

<ShapesGraph#NamePropertyShape> a shacl:PropertyShape;
  shacl:path dwc:scientificName; shacl:nodeKind shacl:Literal;
  shacl:minCount 1; shacl:maxCount 1.

<ShapesGraph#PhotographShape> a shacl:NodeShape;
  ...
```

Listing 4: Snippet of the shapes graph of SPARQL micro-service $S_{\mu f}$.

micro-services. Each step of the processing (selection, matchmaking, query rewriting) is performed using SPARQL queries that involve the SD documents, the shapes graphs and the input query.

3.3 Invocation

To process an incoming SPARQL query, a SPARQL micro-service needs to extract the input arguments from the query graph pattern. For instance, when a client invokes $S_{\mu f}$ with query Q_1 (Listing 1), $S_{\mu f}$ must extract the object of property `dwc:scientificName` (*Delphinus delphis*) to perform the subsequent invocation of Flickr's Web API. This involves reasoning simultaneously on the query graph pattern, the SD document that describes the arguments mappings, and optionally the shapes graph if the mappings refer to property shapes.

Since a SPARQL graph pattern is not represented in RDF, we first translate the incoming query into its SPIN representation [15] that we load into the local triple store as a temporary graph. A major advantage of this approach is that extracting the input arguments can be carried out declaratively within a single SPARQL query rather than in custom code. This query is shown in Listing 5. The first member of the UNION clause (lines 4-11) matches the case where arguments are denoted with `hydra:property`: it retrieves the object of `hydra:property` (line 8), i.e. `dwc:scientificName`, and looks for it in the SPARQL query SPIN graph (line 11). By contrast, the second member (lines 15-34) matches the case where arguments are denoted with a property shape.

Once the arguments have been extracted, the rest of the SPARQL query evaluation is performed as illustrated in section 2.

Implementation. To implement this solution, we deployed Corese [7], an in-memory triple store, as the SPARQL engine underlying SPARQL micro-services. Corese implements the SPARQL Template Transformation Language (STTL) [5] and comes with a built-in STTL SPARQL-to-SPIN transformation. For greater flexibility, our implementation allows passing arguments with VALUES or FILTER clauses, which entails a substantially more complicated query than the one depicted in Listing 5. In particular, it leverages

```
SELECT DISTINCT ?name ?predicate ?value
WHERE {
 {
  # Predicate given with hydra:property
  [] a sd:Service;
   dct:source [ schema:potentialAction [ hydra:mapping [
   hydra:variable ?name;
   hydra:property ?predicate;
  ]]].

  [] sp:predicate ?predicate; sp:object ?value.
 }
 UNION
 {
  # Predicate given through a property shape
  [] a sd:Service;
   dct:source [ schema:potentialAction [
     hydra:mapping [
       hydra:variable ?name;
       shacl:sourceShape ?propShape;
  ]]].

  ?nodeShape a shacl:NodeShape; shacl:property ?propShape.
  ?propShape a shacl:PropertyShape; shacl:path ?predicate.
  OPTIONAL {
   ?nodeShape shacl:property [
    shacl:path rdf:type; shacl:hasValue ?class
   ]
  }

  [] sp:subject ?subject; sp:predicate ?predicate; sp:object ?value.
  OPTIONAL {
   [] sp:subject ?subject; sp:predicate rdf:type; sp:object ?class.
  }
 }
}
```

Listing 5: Extraction of input argument predicates and values (variables `?predicate` and `?value`) from a SPARQL query submitted to a SPARQL micro-service. The default dataset contains the SPARQL query SPIN graph, the shapes graph and the SPARQL SD graph.

the LDScript [6] SPARQL extension to define functions able to parse the nested RDF lists entailed by the VALUES clause[15].

From a more general perspective, the approach we propose considers the service as a coherent, self-contained, reflexive system where RDF and SPARQL are used internally for the service self-description and configuration, at run-time for the query processing, and as the service external interface.

4 WEB-SCALE DISCOVERY OF SPARQL MICRO-SERVICES

In section 1, we suggested that Web search engines can play a key role in enabling the automatic discovery and querying of data sources at Web-scale. Applied to our context, this means that SPARQL micro-services should be published along with a dedicated Web page to be indexed by search engines. Furthermore, major search engines now recommend the inclusion of markup data in Web pages to enhance indexing and consequently yield more accurate results. Therefore, to spur Web-scale discovery while avoiding redundant work, we propose that SPARQL micro-services dynamically transform their service description into Web pages that embed rich markup data meant for search engines. Following content negotiation principles, the micro-service URL dereferences to this Web page if it is looked up by a Web browser, while it dereferences

[15]The complete query is available at https://github.com/frmichel/sparql-micro-service/tree/0.3.1/src/sparqlms/resources/read_input_from_gp.sparql.

to the SPARQL SD document when requested with appropriate RDF media types.

To standardize such markup data, Google, Yahoo, Bing and Yandex support the Schema.org community project that has become a de-facto standard. In particular, Google's recently launched Dataset Search service[16] exploits Schema.org's `Dataset` term[17] as well as equivalent terms from the DCAT W3C recommendation [21]. A Schema.org `Dataset` consists of a set of distributions represented by means of the `DataDownload` object that, unfortunately, is not suited to depict API resources such as SPARQL endpoints. Ongoing discussions are held within the Schema.org community regarding how to annotate a `Dataset` with the interfaces that allow access it[18]. Until a consensus be eventually adopted, a common workaround implemented by the CKAN data portal[19] is to associate to the `DataDownload` object the encoding format "api/sparql". Although semantically questionable ("api/sparql" is not a standard IANA media type[20]), this practice is a trade-off between the need for valid semantic description and the need for effective Web-scale discovery means. Furthermore, given the popularity of CKAN for hosting data portals, this practice tends to spread out.

In the context of SPARQL micro-services, we mitigate this issue with a twofold approach. On the one hand, we comply with the `DataDownload` + "api/sparql" encoding format practice to ensure maximum discoverability. On the other hand, we embed additional DCAT `Dataset` and `Distribution` objects conveying similar information in a more semantically formal manner. Both ways are depicted in Listing 6, lines 24-30 and 38-49 respectively.

Results. The combination of standard content negotiation, semantic Web standards and current Linked Data practices fuels a human-friendly documentation and testing interface on one side and a machine-readable Linked Data description on another side. Furthermore, this combined use pushes "RDF in HTML" descriptions to Web crawlers and indexes in such a way that the described services can be effectively discovered and called by both humans and machines as if they were just another data source.

As an illustration, at the time of writing, the example service $S_{\mu f}$ can be discovered in Google Dataset Search using the keywords "biodiversity" and "photography". Figure 3 shows a snapshot of the result page. Notice that the available download format is appropriately set to SPARQL. Furthermore, adding keyword "sparql" returns the micro-service as the first result in the result page.

Implementation. The Web page generation is performed using the technologies already introduced in section 3.3. An STTL transformation[21] instantiates HTML templates with elements from the SPARQL SD document. The embedded markup data (exemplified in Listing 6) is generated by a SPARQL CONSTRUCT query whose result is passed to a generic built-in STTL transformation that serializes RDF data in JSON-LD. All these transformations are independent of any service and domain. The whole process happens at run-time upon look-up of the micro-service URL. A snapshot of the Web page generated by service $S_{\mu f}$ is displayed in Figure 4,

```
37 {
38  "@context": [
39    "http://schema.org",
40    { "dcat": "http://www.w3.org/ns/dcat#",
41      "frmt": "http://www.w3.org/ns/formats/" }
42  ],
43
44  "@type": "Dataset",
45  "identifier":
46   "http://sms.i3s.unice.fr/sparq-ms/flickr/getPhotosByTaxon_sd/",
47  "name": "Search for Flickr photos by scientific name",
48  "description": "...",
49  "publisher": {
50   "@type": "Organization",
51   "name": "Universite Cote d'Azur, CNRS, Inria, I3S",
52  },
53  "keywords": [ "biodiversity", "lifesciences", "photography" ],
54  "isBasedOn": {
55   "@type": "CreativeWork",
56   "@id": "https://www.flickr.com/services/api/",
57   "name": "Flickr Web API"
58  },
59
60  "distribution": {
61   "@type": "DataDownload",
62   "contentUrl":
63     "http://sms.i3s.unice.fr/sparq-ms/flickr/getPhotosByTaxon_sd/",
64   "name": "SPARQL endpoint",
65   "description": "SPARQL micro-service endpoint",
66   "encodingFormat": "api/sparql"
67  },
68
69  "additionalType": [
70   "dcat:Dataset",
71   "http://ns.inria.fr/sparql-micro-service#Service"
72  ],
73
74  "dcat:distribution": {
75   "@type": "dcat:Distribution",
76   "name": "SPARQL endpoint",
77   "description": "SPARQL micro-service endpoint",
78   "dact:accessUrl":
79     "http://sms.i3s.unice.fr/sparq-ms/flickr/getPhotosByTaxon_sd/",
80   "dcat:mediaType": [
81     "frmt:SPARQL_Results_JSON", "frmt:Turtle", "frmt:JSON-LD"
82   ],
83   "http://www.w3.org/ns/shacl#shapesGraph":
84     "http://sms.i3s.unice.fr/sparq-ms/flickr/ \
85        getPhotosByTaxon_sd/ShapesGraph"
86  }
87 }
```

Listing 6: Snippet of the JSON-LD markup data embedded in the automatically generated Web page of SPARQL micro-service $S_{\mu f}$.

and the reader may access this page by pointing a Web browser at http://sms.i3s.unice.fr/sparql-ms/flickr/getPhotosByTaxon_sd/.

5 RELATED WORKS

The work presented in this article addresses two fundamental questions that have been studied under many different perspectives: capturing the functionality of Web services on one side, and automating their discovering and consumption by software agents on the other side.

Works about semantic Web services, whether "big" WSDL-based (e.g. OWL-S [4], SAWSDL [8]) or REST-based (e.g. WADL [13]), have long tackled the question of capturing the functionality of a service through the semantic description of their inputs, outputs and the way they relate to one another. These models, that support automatic discovery, invocation and composition of Web services, usually entail the deployment of complex frameworks requiring advanced skills and tooling. Besides, service discovery is made possible using a centralized repository such as the Universal

[16] Google Dataset Search: http://g.co/datasetsearch
[17] Schema.org Dataset: https://schema.org/Dataset
[18] https://github.com/schemaorg/schemaorg/issues/1423. Accessed Jan. 30th 2019.
[19] The CKAN project: https://ckan.org/
[20] IANA media types: https://www.iana.org/assignments/media-types/
[21] The whole transformation code is available at: https://frama.link/hBDkM7ep.

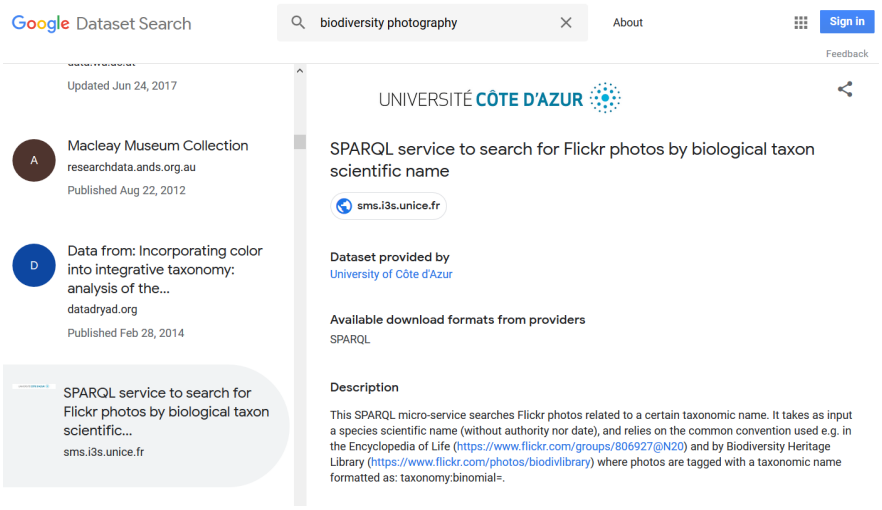


Figure 3: Google Dataset Search result page showing the example SPARQL micro-service $S_{\mu f}$. Clicking on the *sms.i3s.unice.fr* link opens the micro-service Web page depicted in Figure 4.

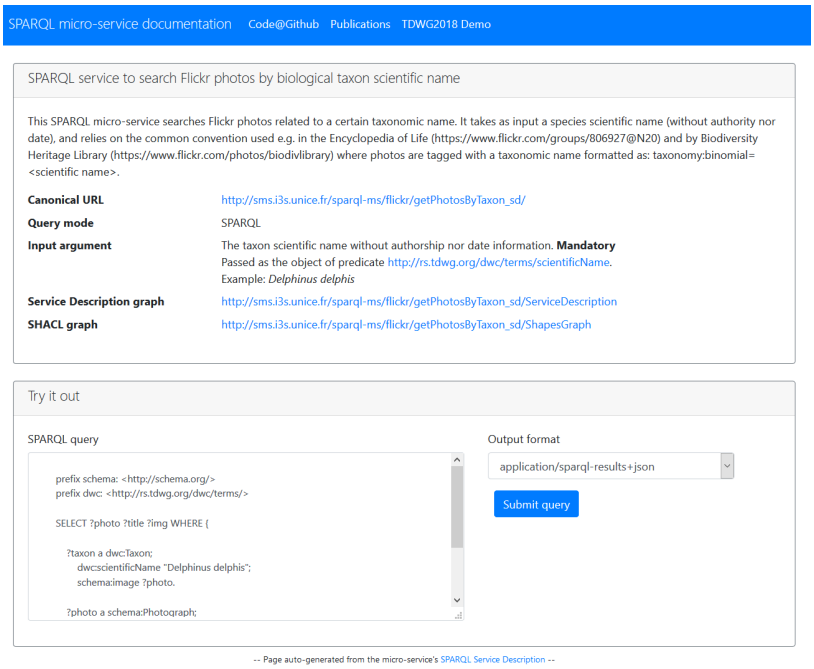


Figure 4: Web page automatically generated from the description of SPARQL micro-service $S_{\mu f}$.

Description Discovery and Integration (UDDI) registry[22]. As a consequence, they are better suited to the controlled environment of companies [20] than the open, loosely constrained environment of the Web that we wish to address.

By contrast, Web APIs are quite simple to deploy and interact with. Still, it is hardly possible to discover and invoke them automatically insofar as they commonly rely on proprietary vocabularies described in Web-based documentation with little concern for semantic interoperability. To fulfill this lack, some initiatives seek to enrich existing human-readable documentation of Web APIs with markup data so as to make it machine-processable. They rely on microformats (e.g. hRESTS [17]) possibly joined to existing service ontologies (e.g. MicroWSMO [9]), or RDFa (e.g. SA-REST [10]). These methods are however more concerned with describing the service interface (operations, parameter types) than its actual functionality. Indeed, the description of the resources manipulated is

[22]UDDI specification: http://uddi.org/pubs/uddi-v3.0.2-20041019.htm

delegated to domain ontologies that provide terms for classes and properties, but often put little constraints on how to use them. By contrast, we harness SHACL specifically to address this lack. SHACL can describe rich constraints on what can be stated, thus making it possible to specify in a comprehensive manner how resources relate to each other.

OpenAPI[23] takes the problem the other way round: it equips a Web APIs with a machine-readable documentation, that, in turn, can be compiled into a Web page. This is closer to our approach, yet, this description remains at a syntactic level essentially enabling the automatic generation of server- and clients-side stubs, very similar to what WSDLs enabled for "heavy" Web services.

Linked REST APIs (LRA) [29] is a framework dedicated to the semantic annotation of Web APIs and the automatic specification of SPARQL query execution plans that invoke these Web APIs. The framework relies on a centralized repository that stores the Web APIs descriptions and offers search services. Several key differences with our work can be pointed out. With SPARQL micro-services, we seek to set up a totally distributed architecture wherein independent service providers may publish SPARQL micro-services that can be discovered using regular Web search engines, rather than a centralized repository. Furthermore, LRA describes a Web API by means of a custom vocabulary and relies on a SPARQL graph pattern to serve as a functional description. To spur large adoption, we instead stick to standard vocabularies, and we use SHACL to describe resources as it allows for more expressiveness than a sheer SPARQL graph pattern.

RESTdesc [32] is a semantic description format for hypermedia APIs. It captures the functional description of APIs in Notation3 [3], a language extending RDF's data model with variables, existential and universal quantifiers, and logical implications. RESTdesc relies on the HTTP mechanisms and RESTful principles for the discovery and invocation of semantically described Web services. Starting from a known URI, an application can follow its nose by resolving links and making sense of Notation3 service descriptions. This is an elegant solution that however requires Notation3 reasoners able to interpret the advanced features of quantification and logical implications. Such reasoners exist but are far less common than SPARQL-based implementations available in many programming languages. Since we seek a solution that can be adopted easily by a large community of independent actors, leveraging more common standards such as regular RDF and SPARQL is probably more promising.

In line with our idea of leveraging Web search engines to discover relevant datasets and query services, SpEnD [34] is a metacrawler designed to discover SPARQL endpoints. It first creates a list of keywords commonly found on Web pages advertising SPARQL endpoints, such as the pages of DataHub. It then looks for these keywords on search engines, explores the result Web pages looking for SPARQL endpoint URLs and looks up these URLs in search for VoID or SPARQL SD documents. This kind of approach is clearly what could be implemented to discover SPARQL micro-services at Web scale. We believe that the usage of well-adopted markup data could help enhance search results and, in this respect, dataset-search services such as Google Dataset Search could be more effective than generic Web search engines.

[23] https://github.com/OAI/OpenAPI-Specification

6 CONCLUSION AND PERSPECTIVES

In this article, we address the problem of enabling automatic discovery and consumption of data sources at Web scale. We suggested that three principles should be considered to pursue this goal: (1) provision rich descriptions of data sources and query services, (2) leverage the power of Web search engines to discover data sources, and (3) rely on simple, well-adopted standards that come with extensive tooling. We applied these principles to the concrete case of SPARQL micro-services that aim at querying Web APIs using SPARQL. The proposed solution considers a SPARQL Service Description (SD) document as the description central point. It links to a SHACL shapes graph describing precisely the resources manipulated by the micro-service. It also connects the resources to the micro-service inputs, thereby coming up with a rich functional description that allows a software agent to decide whether this micro-service can help in carrying out a certain task. To enable accurate discovery using common Web crawlers, the SD document can be dynamically transformed into a Web page embedding rich markup data based on Schema.org's Dataset term and the DCAT vocabulary.

From a general perspective, the combination of standard content negotiation, semantic Web standards and Linked Data practices fuels a human-friendly documentation and machine-readable description that make it possible for humans and machines alike to discover and invoke SPARQL micro-services as if they were just another data source.

We showed that our approach is effective as our example SPARQL micro-service can be successfully discovered using the Google Dataset Search engine (as illustrated in Figure 3). From this point on, a framework such as SpEnD (described in section 6) could be extended to accommodate the invocation of SPARQL micro-services. Service composition-based query answering systems could fetch the shapes graphs of candidate SPARQL micro-services, check the compatibility of their inputs and outputs with respect to the query to process, and finally compute and enact valid compositions. In particular, SPARQL query federation is a specific type of Web service composition wherein any piece of data in the federated graphs may play the role of either an input or an output. Existing federated query engines could be extended so as to reason on the description of SPARQL micro-services and come up with query plans that respect SPARQL micro-services' input requirements.

As pointed out in section 4, denoting a SPARQL endpoint using Schema.org terms is still quite unpractical at the moment. From a more general perspective, describing the multiple interfaces that a client may use to access a dataset is an increasingly pressing need. The Schema.org community is currently thinking this through with discussions revolving around the `Dataset`, `WebAPI` and `EntryPoint` terms. Concomitantly, the DCAT community is working out the next version of the W3C DCAT recommendation [11] that defines the generic concept of `DataService` meant to serve dataset distributions. The term is flexible enough to accommodate various types of interfaces, providing notably a contract the interface conforms to

and an out-of-band description that may typically be a SPARQL SD document in our context.

In the current state of our work, SHACL graphs are used as a specification of the graphs that a SPARQL micro-service can generate. We can think of two interesting leads for future works in this respect. Firstly, once a shapes graph is published with its own dereferenceable URI, it can be reused by SPARQL micro-services providers, thereby sparing time and making it possible to share common practices. A second lead could be to consider SHACL as a way for a client to request responses in a certain shape. This would amount to some sort of extended content negotiation where a client could express that it would prefer a response not only favoring a vocabulary over another, but also describing resources and their relationships according to a certain shape, as much as possible.

Finally, whether our approach succeeds in reaching principle (3) (rely on well-adopted standards) is debatable and possibly a matter of perspective and community. Some people contend that Semantic Web standards are not likely to be largely adopted by Web developers [18] due to the perceived complexity of RDF and SPARQL, as compared to RESTful APIs for instance. Besides, SHACL is a rich language, yet perhaps too rich to gain large adoption. In the end however, we do believe that there will be room for different types of interfaces, suited to different contexts and scenarios. This article primarily intends to propose a research direction, not a ready-to-use solution. And we encourage the interested readers to explore alternative architectural and modeling choices.

REFERENCES

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 2011. Describing Linked Datasets with the VoID Vocabulary. *W3C Recommendation* (2011).

[2] Mohamed Ben Ellefi, Zohra Bellahsene, John G. Breslin, Elena Demidova, Stefan Dietze, Julian Szymański, and Konstantin Todorov. 2018. RDF Dataset Profiling – a Survey of Features, Methods, Vocabularies and Applications. *Semantic Web – Interoperability, Usability, Applicability* 9, 5 (2018), 677–705. https://doi.org/10.3233/SW-180294

[3] Tim Berners-Lee and Dan Connolly. 2011. Notation3 (N3): A Readable RDF Syntax. *W3C Team Submission* (March 2011).

[4] Mark Burstein, Jerry Hobbs, Ora Lassila, Drew McDermott, Sheila McIlraith, Srini Narayanan, Massimo Paolucci, Bijan Parsia, Terry Payne, Evren Sirin, Naveen Srinivasan, and Katia Sycara. 2004. OWL-S: Semantic Markup for Web Services. *W3C Member Submission* (2004).

[5] Olivier Corby and Catherine Faron-Zucker. 2015. STTL: A SPARQL-Based Transformation Language for RDF. In *11th International Conference on Web Information Systems and Technologies (ISWC)*. Lison, Portugal.

[6] Olivier Corby, Catherine Faron-Zucker, and Fabien Gandon. 2017. LDScript: A Linked Data Script Language. In *Proceedings of the 16th International Semantic Web Conference (ISWC)*. Springer, Vienna, Austria, 208–224.

[7] Olivier Corby and Catherine Faron Faron-Zucker. 2010. The KGRAM Abstract Machine for Knowledge Graph Querying. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 338–341.

[8] Joel Farrell and Holger Lausen (Eds.). 2007. Semantic Annotations for WSDL and XML Schema. *W3C Recommendation* (2007).

[9] Florian Fischer and Barry Norton. 2009. *MicroWSMO v2 – Defining the Second Version of MicroWSMO as a Systematic Approach for Rich Tagging*. Technical Report.

[10] Karthik Gomadam, Ajith Ranabahu, and Amit Sheth. 2010. SA-REST: Semantic Annotation of Web Resources. *W3C Member Submission* (2010).

[11] Alejandra Gonzalez Beltran, Dave Browning, Simon Cox, and Peter Winstanley. 2019. Data Catalog Vocabulary (DCAT) - revised edition. *W3C Editor's Draft 01 February 2019* (2019). https://w3c.github.io/dxwg/dcat/ Accessed: 2019-02-01.

[12] R. V. Guha, Dan Brickley, and Steve MacBeth. 2015. Schema.Org: Evolution of Structured Data on the Web. *ACM Queue - Strutured Data* 13, 9 (2015), 10:10–10:37. https://doi.org/10.1145/2857274.2857276

[13] Marc Hadley. 2009. Web Application Description Language. *W3C Member Submission* (2009).

[14] Steve Harris and Andy Seaborne. 2013. SPARQL 1.1 Query Language. *W3C Recommendation* (2013). http://www.w3.org/TR/2013/REC-sparql11-query-20130321/

[15] Holger Knublauch. 2013. SPIN - SPARQL Syntax. *W3C Member Submission* (2013).

[16] Holger Knublauch and Dimitris Kontokostas (Eds.). 2017. Shapes Constraint Language (SHACL). *W3C Recommendation* (2017).

[17] Jacek Kopecky, Karthik Gomadam, and Tomas Vitvar. 2008. hRESTS: An HTML Microformat for Describing RESTful Web Services. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT)*. 619–625. https://doi.org/10.1109/WIIAT.2008.379

[18] Markus Lanthaler and Christian Gütl. 2011. A Semantic Description Language for RESTful Data Services to Combat Semaphobia. In *Proceedings of the IEEE International Conference on Digital Ecosystems and Technologies*. 47–53. https://doi.org/10.1109/DEST.2011.5936597

[19] Markus Lanthaler and Christian Gütl. 2013. Hydra: A Vocabulary for Hypermedia-Driven Web APIs. In *Proceedings of the 6th Workshop on Linked Data on the Web (LDOW2013)*. CEUR-WS.

[20] Moussa Lo and Fabien Gandon. 2007. Semantic Web Services in Corporate Memories. In *Proceedings of the Second International Conference on Internet and Web Applications and Services*. https://doi.org/10.1109/ICIW.2007.59

[21] Fadi Maali, John Erickson, and Phil Archer. 2014. Data Catalog Vocabulary (DCAT). *W3C Recommendation* (Jan. 2014).

[22] Franck Michel, Catherine Faron-Zucker, and Fabien Gandon. 2018. SPARQL Micro-Services: Lightweight Integration of Web APIs and Linked Data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2018)*, Vol. 2073. CEUR, Lyon, France, 10.

[23] Franck Michel, Catherine Zucker, Olivier Gargominy, and Fabien Gandon. 2018. Integration of Web APIs and Linked Data Using SPARQL Micro-Services—Application to Biodiversity Use Cases. *Information* 9, 12 (Dec. 2018), 310. https://doi.org/10.3390/info9120310

[24] Sam Newman. 2015. *Building Microservices*. O'Reilly Media.

[25] Cesare Pautasso, Olaf Zimmermann, and Frank Leymann. 2008. RESTful Web Services vs. "Big" Web Services: Making the Right Architectural Decision. In *Proceedings of the 17th International World Wide Web Conference (WWW2008)*. Beijnig, China, 805–814. https://doi.org/10.1145/1367497.1367606

[26] Emmanuel Pietriga, Hande Gözükan, Caroline Appert, Marie Destandau, Šejla Čebirić, François Goasdoué, and Ioana Manolescu. 2018. Browsing Linked Data Catalogs with LODAtlas. In *Proceedings of the 17th International Semantic Web Conference (ISWC)*. Monterey, USA, 17.

[27] Mike Ralphson and Ivan Goncharov. 2017. WADG0001 WebAPI type extension. *Draft Community Group Report* (2017). https://webapi-discovery.github.io/rfcs/rfc0001.html

[28] Andy Seaborne. 2013. SPARQL 1.1 Query Results JSON Format. *W3C Recommendation* (2013).

[29] Diego Serrano, Eleni Stroulia, Diana Lau, and Tinny Ng. 2017. Linked REST APIs: A Middleware for Semantic REST API Integration. In *Proceedings of the IEEE International Conference on Web Services (ICWS)*. IEEE, Honolulu, HI, USA, 138–145. https://doi.org/10.1109/ICWS.2017.26

[30] Manu Sporny, Dave Longly, Gregg Kellog, Markus Lanthaler, and Niklas Lindström. 2014. JSON-LD 1.0. A JSON-based Serialization for Linked Data. *W3C Recommendation* (2014). https://www.w3.org/TR/2014/REC-json-ld-20140116/

[31] Pierre-Yves Vandenbussche, Jürgen Umbrich, Luca Matteis, Aidan Hogan, and Carlos Buil-Aranda. 2017. SPARQLES: Monitoring Public SPARQL Endpoints. *Semantic Web* 8, 6 (Aug. 2017), 1049–1065. https://doi.org/10.3233/SW-170254

[32] Ruben Verborgh, Thomas Steiner, Davy Van Deursen, Jos De Roo, Rik Van de Walle, and Joaquim Gabarro Vallés. 2011. Description and Interaction of RESTful Services for Automatic Discovery and Execution. In *Proceedings of the International Workshop on Advanced Future Multimedia Services*. Future Technology Research Association International (FTRA), Jeju, South Korea.

[33] Gregory Tood Williams (Ed.). 2013. SPARQL 1.1 Service Description. *W3C Recommendation* (2013).

[34] Semih Yumusak, Erdogan Dogdu, Halife Kodaz, Andreas Kamilaris, and Pierre-Yves Vandenbussche. 2017. SpEnD: Linked Data SPARQL Endpoints Discovery Using Search Engines. *IEICE Transactions on Information and Systems* E100.D, 4 (2017), 758–767. https://doi.org/10.1587/transinf.2016DAP0025

LocWeb2019 Chairs' Welcome

We are excited to welcome you to the *9th International Workshop on Location and the Web (LocWeb2019)* at *WWW 2019*. LocWeb 2019 will continue the workshop series at the intersection of location-based services and Web architecture. It focuses on Web-scale services and systems facilitating location-aware information access as well as on Spatial Social Behavior Analytics on the Web as part of social computing.

LocWeb addresses location as a cross-cutting issue in web research and technology that connects the online world to the physical spatial world. It examines location aspects in the domains of social computing, search, analytics, mobility, apps, services, standards, and systems. The aspect of social computing will have a particular focus on spatio-social networks, i.e., social network users' behaviors, activities, interactions, news propagation, and social influence learning models in the context of available spatial or location information. We are particularly interested in new tools and techniques for the modeling and analysis of dynamic spatio-social networks, as well as understanding of the disconnect between activities in the online and physical worlds and methods and techniques to bridge this gap.

We received 10 submissions from all over the world and could accept 4 of them to build an engaging workshop. Review and acceptance statistics are as follows:

Submission Type	Reviewed	Accepted	
Full Technical Papers	7	4	57%
Short Poster Presentations	3	0	0%
Demos, Tutorials, or Others	10	4	40%

We encourage WWW attendees to attend the LocWeb presentations, keynotes, and discussion sessions. The detailed programme of the half-day workshop will be available on the workshop website: https://dhere.de/locweb/locweb2019/

Putting LocWeb2019 together was a team effort. We thank the authors for their work, the programme committee for substantial reviews under tight deadlines, and The Web Conference for hosting us.

Dirk Ahlers
NTNU
Norway

Erik Wilde
CA Technologies
Switzerland

Rossano Schifanella
University of Turin
Italy

Jalal S. Alowibdi
University of Jeddah
Saudi-Arabia

THE WEB CONFERENCE

LocWeb2019 Organization

Workshop/Track Chairs:
Dirk Ahlers *(NTNU)*
Erik Wilde *(CA Technologies)*
Rossano Schifanella *(University of Turin)*
Jalal S. Alowibdi *(University of Jeddah)*

Programme Committee:
Arjen de Vries *(Radboud University)*
Bart Thomee *(Google)*
Bruno Martins *(IST and INESC-ID – Instituto Superior Técnico, University of Lisbon)*
Carsten Keßler *(Department of Planning, Aalborg University Copenhagen)*
Chien-Wen Shen *(National Central University)*
Clodoveu Davis *(Universidade Federal de Minas Gerais)*
Francisco López Pellicer *(University of Zaragoza)*
Leonidas Anthopoulos *(University of Applied Science (TEI) of Larissa, Greece)*
Lisette Espín Noboa *(GESIS – Leibniz-Institute for the Social Sciences, University of Koblenz-Landau)*
Luca Maria Aiello *(Nokia Bell Labs)*
Marcos Oliveira *(GESIS – Leibniz Institute for the Social Sciences)*
Muhammad Usman Ilyas *(University of Jeddah)*
Rainer Simon *(AIT – Austrian Institute of Technology)*
Salem Alelyani *(King Khalid University)*
Torsten Suel *(New York University)*
Yana Volkovich *(Xandr)*

Indoor Towers, DPIs, and More People in Parks at Night: New Trends in Mobile Phone Location Research

Leo Ferres
lferres@udd.cl
Data Science Institute
Faculty of Engineering, Universidad del Desarrollo
Telefónica R&D
Santiago, Chile

ABSTRACT

In this keynote, I will present some results we have obtained using location data from mobile phones interacting with information (news) websites, mobile apps like Pokemon GO and Twitter, and large physical spaces like shopping malls. I will draw some conclusions and generally discuss about the properties of mobile phone data for location-based research to finally close with some remarks about privacy and data security.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing*.

KEYWORDS

Mobile phones, CDR and XDR datasets, deep packet inspection, data science for social good, location data mining

ACM Reference Format:
Leo Ferres. 2019. Indoor Towers, DPIs, and More People in Parks at Night: New Trends in Mobile Phone Location Research. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316539

1 INTRODUCTION

We're more connected than ever. In part, this is due to the high availability (in both developed and developing countries) of relatively cheap smart phones that we carry with us all the time. In fact, in 2018, more than 52% of the whole website traffic was generated by mobile phones. Unlike desktops or laptops, one defining characteristic of mobile phones is that they are always geo-located, either by GPS, triangulation, or antenna connections. Thus, mobile phone data sets, either Call Detail Records (CDRs) or Data Detail Records (XDRs) constitute a potential treasure trove of information about what people do in the physical world, not only when interacting with it, but also when accessing information online.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316539

2 DATA AND STUDIES

Through our association to Telefónica Chile R&D department, we have been privileged in our access to mobile phone data sets to study socially relevant issues. In this talk, I will present results from several studies done using ecologically-valid, very large data sets of billions of mobile phone usage data (CDRs, XDRs, and a bit of an even lower level of analyses, deep packet inspection) and the towers they connect to, drawing conclusions and predicting different kinds of behaviors, from social mixing [1], to news consumption, to gender equality. All these studies were conducted by analyzing web traffic either by proxy to applications like Pokemon Go [3], or effectively through DNS resolution as in the news study or as in the apps study [2]. I will conclude by talking about the coming trends in the field of mobile phone data analysis, its limitations, and spend some time discussing issues of privacy, data security, anonymization and general data responsibility for researchers and the company providing the data.

3 BIO

Leo Ferres is an associate professor of computer science at the Data Science Institute, Universidad del Desarrollo in Santiago, Chile, and a Fellow of Telefónica Research & Development, also in Santiago. He has published widely in high-performance computing, computational social science, using mobile and telephony data for social good, and data mining in general. Together with UNICEF, the ISI Foundation, and The GovLab, he is now investigating issues pertaining to gender and gender gaps, social inclusion at the city scale, and news diffusion using mobile phone data.

ACKNOWLEDGMENTS

Grateful to Movistar-Telefónica Chile, CORFO 13CEE2-21592, Conicyt PAI Networks (REDES170151), and Project PLU180009, 2018.

REFERENCES

[1] Mariano G. Beiró, Loreto Bravo, Diego Caro, Ciro Cattuto, Leo Ferres, and Eduardo Graells-Garrido. 2018. Shopping mall attraction and social mixing at a city scale. *EPJ Data Science* 7, 1 (31 Aug 2018), 28. https://doi.org/10.1140/epjds/s13688-018-0157-5

[2] Eduardo Graells-Garrido, Diego Caro, Omar Miranda, Rossano Schifanella, and Oscar F. Peredo. 2018. The WWW (and an H) of Mobile Application Usage in the City: The What, Where, When, and How. In *Companion Proceedings of the The Web Conference 2018 (WWW '18).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1221–1229. https://doi.org/10.1145/3184558.3191561

[3] Eduardo Graells-Garrido, Leo Ferres, Diego Caro, and Loreto Bravo. 2017. The effect of Pokémon Go on the pulse of the city: a natural experiment. *EPJ Data Science* 6, 1 (15 Sep 2017), 23. https://doi.org/10.1140/epjds/s13688-017-0119-3

Location Embeddings for Next Trip Recommendation

Amine Dadoun
EURECOM, Sophia Antipolis, France
Amadeus SAS, Biot, France
amine.dadoun@eurecom.fr

Raphaël Troncy
EURECOM, Sophia Antipolis, France
raphael.troncy@eurecom.fr

Olivier Ratier
Amadeus SAS, Biot, France
olivier.ratier@amadeus.com

Riccardo Petitti
Amadeus SAS, Biot, France
riccardo.petitti@amadeus.com

ABSTRACT

The amount of information available in social media and specialized blogs has become useful for a user to plan a trip. However, the user is quickly overwhelmed by the list of possibilities offered to him, making his search complex and time-consuming. Recommender systems aim to provide personalized suggestions to users by leveraging different type of information, thus assisting them in their decision-making process. Recently, the use of neural networks and knowledge graphs have proven to be efficient for items recommendation. In our work, we propose an approach that leverages contextual, collaborative and content information in order to recommend personalized destinations to travelers. We compare our approach with a set of state of the art collaborative filtering methods and deep learning based recommender systems.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → Neural Networks.

KEYWORDS

Recommender Systems, Neural Networks, Knowledge Graph, Embeddings, Tourism

ACM Reference Format:
Amine Dadoun, Raphaël Troncy, Olivier Ratier, and Riccardo Petitti. 2019. Location Embeddings for Next Trip Recommendation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316535

1 INTRODUCTION

Traveling is no longer considered to be just a need and it has clearly become a desire. Users became exposed to many inspirational tourism posts and advertisements in social media, travel forums, travel agencies and airline websites. Although inspirational, many of these posts might not fit a particular user's profile and, thus, they may not be relevant to him. In recent years, destination recommender systems (DRSs) have been proposed to suggest a ranked list of destinations, sometimes composed of sights, events and cities to visit, based on information provided by the user [10, 13, 27]. Recommender systems can also take contextual information into account, for example, by leveraging event-based social networks data [15]. Location-based Social Networks (LBSNs) allow users to publicly or privately share their position by performing a check-in when visiting a certain venue or Point-of-Interest (POI). Leveraging this data enables to first know what a city is best characterized by (restaurants, sport events, museums, parks, etc.), and then to identify the user's interests [17]. Finally, a user profile can be enriched with his booking history.

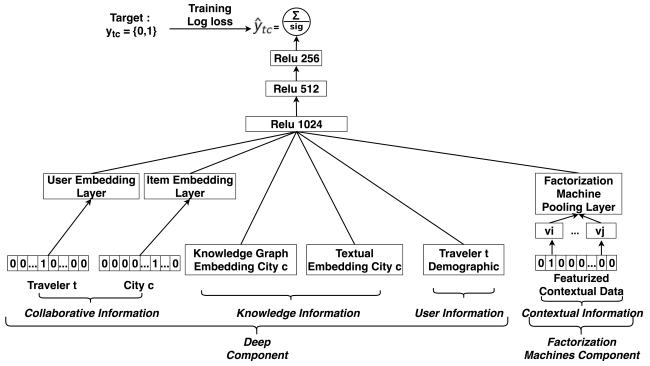


Figure 1: Deep Knowledge Factorization Machines architecture for next trip destination.

Recent works have illustrated the effectiveness of using knowledge graph embeddings (KGE) [19, 24, 28] and neural networks [2, 7, 8] for item recommendation. In this paper, we propose a Deep Knowledge Factorization Machines (DKFM) architecture to recommend destinations. Our approach relies on learning *i)* a representation of cities using different data sources including Wikipedia and LBSN, *ii)* the long-term user's behavior using his booking history and *iii)* a representation of the context associated with each past trips. More specifically, we combine textual embeddings representing cities based on their wikipedia content description with knowledge graph embeddings that represent cities' characteristics in terms of venue check-ins made by users in LBSNs. We also combine two existing deep learning based recommender systems [2, 7] to build our so-called DKFM architecture (Figure 1) which takes as input content, collaborative and contextual information related to user bookings.

The main contributions of our work are:

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316535

(i) We leverage three types of information (collaborative, content and contextual) in order to recommend a next trip.
(ii) We present a deep neural network model that takes as input these three different types of information.
(iii) We perform an empirical comparison on a real-world dataset of our DKFM architecture with state-of-the-art collaborative filtering methods and deep learning based recommender systems.

The rest of the paper is organized as follows. Section 2 provides a literature review of the related work. Section 3 introduces some preliminaries concepts, the recommendation problem and the dataset we use for the experiments. In Section 4, we present the approach to build the DKFM architecture. Section 5 presents the experiments carried out to show the effectiveness of our model. Finally, in Section 6, we provide some conclusions and we discuss some future work.

2 RELATED WORK

This section provides a literature review of recommender system in the tourism domain, which has known growing interest from many academic and industrial researchers, trying to tackle several recommendation tasks such as recommending POIs, suggesting sequence of attractions/activities to do or proposing a complete trip plan. This section presents state-of-the-art methods that utilize knowledge graph embeddings and deep learning methods for item recommendation.

Recommender System in Tourism: Tourism includes traveling for business or leisure. It involves complex decision-making from travelers to select destinations, hotels, events, activities, etc. On the other side, travel industry players (e.g. travel agents) are helping travelers to find the most suitable options. Early works have focused on personalized techniques in order to provide recommendation based on user's preferences and interests [22]. More specifically, the idea is to match items in a catalogue of destinations with the user needs, and interests expressed by the offered language. Kiseleva et al. [13] proposed a multi-criteria rating system (MCRS) based on naive bayes approach in order to recommend travel destinations in a hotel booking platform[1]. MCRS are based on which aspects a user liked for a given item, while classical recommendation systems are based on a single rating (e.g. giving a rating for a movie). Wolfgang et al. [26] proposed an approach to generate sequence of POIs when visiting a city based on three user's inputs: start and end point plus interests. However, user preferences or item characteristics are in many cases insufficient to have accurate recommendation. Macedo et al. [15] have proposed to use contextual signals provided by LBSNs such as time or location for events recommendation. In our work, we combine three different types of input: the traveler-destination interaction which is represented by a booking, the context of each booking, and the content that characterizes each destination, to suggest destinations to travelers.

Knowledge Graph Embeddings for Items Recommendation: A Knowledge graph embedding is a representation of a knowledge graph's component into continuous vector space. The idea is to ease the manipulation of graph components (entities, relations) for prediction tasks such as entity classification, link prediction or recommender systems [28]. A survey of approaches and applications for knowledge graph embeddings was done by Wang et al. [25]. Two main approaches exist in order to learn knowledge graph embeddings from a KG: translational distance models where the goal is to minimize the distance between neighbors entities in the graph; semantic matching models which are based on the semantics of the graph components compute a similarity score that measures the semantic similarity between each entity in the graph. In order to enrich the collaborative information represented by the user-item interaction with additional information, knowledge graphs have been used to provide knowledge about items and/or users to enhance the performance of the recommendation. In [28], the authors used a knowledge base containing different external resources: textual information, visual information and structural information (knowledge graph embedding) to enrich the user implicit feedback for items recommendation. In [18], the authors used the concept of property-specific knowledge graph to learn embeddings based on node2vec [5] of each subgraph and then used a ranking function to provide items recommendation. In our work, we use a knowledge graph from LBSN's check-ins in order to learn knowledge graph embeddings that represent cities.

Deep learning based Recommender System: In the recent years, deep learning has demonstrated its effectiveness when applied to information retrieval and recommender system. In [4], the authors used a multilayer perceptron that takes as input the (user, item) interaction and learn user and item embeddings. In [8], the authors combined a multilayer perceptron with a generalization of matrix factorization represented by a single layer perceptron. In [2], the authors proposed wide and deep learning model for app recommendation[2]. The wide learning component is a single layer perceptron which enables to capture memorization and the deep learning component is a multilayer perceptron which enables to capture generalization. In [6], the authors combined factorization machines and multilayer perceptron. The idea is to model the high-order feature interactions via multilayer perceptron and low-order interactions with Factorization Machine [20]. In [7], the authors proposed to use a pooling layer that computes the first order feature interaction term in factorization machines formula [20], then feed the obtained vectors in a multilayer perceptron. Other neural network architectures have been used for recommendation, such as Recurrent Neural Networks (RNNs) (e.g. session-based recommendation) or Convolutional Neural Networks (CNNs) used for example to capture images representation in order to enrich item representation.

Our work focuses on feed-forward neural networks. We combine two existing deep learning based recommender systems [2, 8] to build our recommender system.

3 DATA & PROBLEM FORMULATION

In this section, we first provide definitions of some concepts useful to build our recommendation system. Then, we present the next trip recommendation task. Finally, we present the dataset used to address the next trip recommendation.

[1]Booking.com: https://www.booking.com

[2]Google Play Store: https://play.google.com/store

3.1 Preliminaries

We define a Knowledge Graph similarly to what has been done in [18].

Definition 1. A knowledge graph is a set $K = (E, R, O)$, where E is the set of entities, $R \subset E \times \Gamma \times E$ is a set of typed relations between entities and O is an ontology.

In recommender system realm, there are two different types of feedback: explicit feedback where the user gives a rating on how he liked the item or not, and the implicit feedback where we know only the interest of a user for an item. Concretely, in our case, the implicit feedback denotes the fact that a traveler t visited a destination d.

Definition 2. Given a matrix $\mathbf{T} \in \mathbb{R}^{n \times m}$, where t_{ij} is the number of times the traveler i traveled to destination j, n the number of travelers and m the number of different destinations. We define the traveler binary feedback matrix $\mathbf{R} \in \mathbb{R}^{n \times m}$ as follows:

$$r_{ij} = \begin{cases} 1 & \text{if } t_{ij} > 0 \\ 0 & \text{otherwise .} \end{cases} \quad (1)$$

3.2 Problem Formulation

Definition 3. Given a traveler, his demographics information (age, nationality, etc.), his historical bookings and the contextual data related to those bookings (day of week, number of passengers, stay duration, etc.), we aim to recommend to this traveler a ranked list of destinations he would like to go to. A destination is represented by a city that has an airport. The Figure 2 illustrates the recommendation task we want to tackle.

Figure 2: The recommendation task is to predict a ranked list of next destinations for a traveler given his historical bookings.

3.3 Data Preprocessing

Experiments were conducted on a real-world dataset of bookings from the so-called CEM[3] database, an Amadeus database containing bookings over a dozen of airlines. Each booking is stored using Personal Name Records (PNR) information. The PNR is created at reservation time by the airline and contains information about the purchased ticket (travel itinerary, traveler demographics information, payment information, ancillary services bought with the air ticket, etc.). The considered dataset contains 4.8 Million bookings for 814.919 unique travelers[4].

[3]CEM: Customer Experience Management

[4]Statistics of the pre-processed dataset are given in Tables 3 & 4

Table 1: Features used for business/leisure classification

Feature Name	Type	Range
Number Passenger	Numerical	{1..9}
Stay Duration	Numerical	[0,99]
Saturday Stay	Binary	{0,1}
Purchase Anticipation	Numerical	{0..364}
Age	Numerical	{0..99}
Gender	Categorical	{Female, Male, Unknown}

Table 2: Business/Leisure classification performance

Metric	Score
Accuracy	0.87
Precision	0.87
Recall	0.91

Customer segmentation model: The approach to recommend destinations to business/leisure travelers is expected to be different as explained in [3]. In our work, we focus only on recommending travels for leisure purpose. Hence, we need to split the bookings into business/leisure segments. Given an historical dataset of bookings that are already labeled into business/leisure travels, we build a Random Forest based classifier in order to classify our bookings. Table 1 shows the dataset's features used for this classification task. In order to train our classifier, we use an existing dataset that contains bookings already labeled (Business/Leisure). This dataset was collected from travel agencies from February 2014 to February 2017 and contains 122,242 bookings (60% leisure). The classifier (Random Forest) was tuned using grid-search algorithm over the following hyper-parameters: maximum tree depth taken in {5, 8, 10}, maximum used features taken in {0.6, 0.65, 0.7, 0.75}, minimum samples in leaf taken in {1, 2}, number of trees taken in {100, 150, 200}. Finally, to evaluate our classifier, we used a 10 Fold Cross Validation method by splitting our dataset into training and test sets (90 % training, 10 % test set) and compute the accuracy, precision and recall metrics for the tuned classifier. We also compute the importance of each feature for the classification based on the relative information gain of each feature; The number of passenger is the most important feature for this classification task.

The classifier was then used to classify bookings of the considered dataset into Business/Leisure bookings. We keep only leisure bookings for our work. We obtained 2 Million bookings made for 629.156 unique travelers, which represent 42% of the whole dataset.

Data Filtering for Recommendation: Despite the huge amount of available bookings that could be used to build the recommendation system, we have a very sparse traveler feedback matrix. The sparsity of the traveler binary feedback matrix $\mathbf{R} \in \mathbb{R}^{n \times m}$ is defined as follows:

$$\rho(\mathbf{R}) = 1 - \frac{\#interactions}{m \times n} \quad (2)$$

where, n is the number of users and m is the number of different destinations.

Table 3: Statistics of the preprocessed dataset

#Feedbacks	#Interactions	#Cities	#Travelers	Sparsity
304019	152547	119	26019	95%

Table 4: Statistics of the preprocessed dataset

Variable	Min	Max	Std	Mean	Median
#Visiting same city	1	354	3.34	2	1
#Travelers per city	20	19496	2452	1282	293
#Cities per traveler	5	37	1.49	5.86	5

In fact, for this dataset, the sparsity is 99.6%. Moreover, more than 65% of the travelers have traveled only two times. Similarly to [8, 9], in order to cope with this issue, we keep only travelers that have at least **5** different destinations in their history, and destinations that were visited at least **20** times. After applying this last filter, we obtained a dataset containing 304.019 booking for 26.019 unique travelers. It is important to note that an interaction represents the fact that a traveller has been to a destination at least once, which is different from feedback, which represents the number of times a traveller has been to that destination. Tables 3 & 4 represent statistics of the pre-processed dataset.

4 APPROACH

Our approach is to leverage data from different sources to improve the recommendation by enriching implicit interaction between traveler and destination with external knowledge. Moreover, the context in travel is an important factor to consider when doing recommendation, thus, we also add contextual information related to a given booking. We combine a deep component which is a multilayer perceptron that takes as input the implicit interaction and the content information, with a factorization machines component that takes as input contextual data. The deep and factorization machines components were combined by concatenating:

- Traveler and city embeddings (the destination is represented by a city);
- Textual and knowledge graph city embeddings;
- User demographics information;
- Contextual feature vectors computed by the pooling operation (see Section 4.4).

The concatenated vectors are fed into a multilayer perceptron. The two components are jointly trained using backpropagation algorithm to learn the weights of the deep and factorization machines components, and also the traveler and city embeddings. In this section, we first present how the destinations are enriched with external knowledge resources, and present the different components of our model. Finally, we present how we combine the two existing deep learning based recommender systems to build our model.

4.1 Textual Embedding

In this subsection, the goal is to explore the different ways to build a textual representation of the cities based on the content of their Wikipedia pages. The first step is to retrieve all the Wikipedia pages of the 119 cities covered in our dataset. To do so, we used the Wikipedia Python API[5]. Once all documents describing these cities have been retrieved, we need to define a method to construct an embedding of each document. In the recent years, many competing algorithms were proposed to learn sentence or document representation. In [12], the authors proposed to learn unsupervised sentence embeddings based on RNN encoder-decoder which is trained to reconstruct the surrounding sentences from the current sentence similarly to what is done in skip-gram model for word embeddings. In [14], the authors proposed a faster way to learn unsupervised sentence representations by reformulating the problem as a classification task, where the classifier has to choose the right next sentence among a set of possibilities. While these approaches have shown good performance, simple baseline like averaging pre-trained word embeddings give also strong results. We propose to encode a sentence in a weighted sum of word vectors, where the weight of each word vector corresponds to the term frequency-inverse document frequency (TF-IDF) of the word. We used the fastText pre-trained word vectors [16]. The word vectors were trained using Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset.

4.2 Knowledge Graph Embedding

Similarly to the previous subsection, we explore the different methods to construct a city representation based on knowledge graph. First, we need to build a knowledge graph that contains characteristics of cities. In [17], the authors used LBSNs to build users' trails. A trail is a succession of check-ins made by a user in venues. Each venue is categorized using Schema.org[6] (Restaurant, Civic structure, etc.) and the Foursquare category which is more detailed (Italian restaurant, Indian restaurant, etc.). The authors released the Semantic Trails Datasets[7] which contain two datasets. The first dataset contains check-ins collected in the temporal interval going from 03-04-2012 to 16-09-2013, while the second one contains check-ins collected going from 03-10-2017 to 19-10-2018. To build the Semantic Trail Knowledge Graph in Figure 3, we used both datasets. The knowledge graph represents the interaction user-venue, through the property 'visiting' as well as the relations between the venue and the other entities, namely: category, schema and city.

In [19], the authors presented an empirical comparison of translational distance models for items recommendation. The results have shown that TransE [1], the model with the least parameters in comparison with other translational distance models [25] obtained the best scores over a set of metrics. We propose to use TransE to learn embeddings for the entities and relations in the knowledge graph, and extract the cities embeddings. The idea of TransE algorithm lies in minimizing a distance D, between $h + l$ and t, so that $D(h + l, t) \approx 0$, where the triple (h, l, t) corresponds to (head, relation, tail) entities. Finally, in order to match the cities of our dataset with the cities embeddings obtained by the semantic trails knowledge graph, we used their Wikidata id[8].

[5]Python Wikipedia API: https://pypi.org/project/wikipedia/
[6]Schema.org: https://schema.org
[7]Semantic Trails Datasets: https://figshare.com/articles/Semantic_Trails_Datasets/7429076
[8]Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page

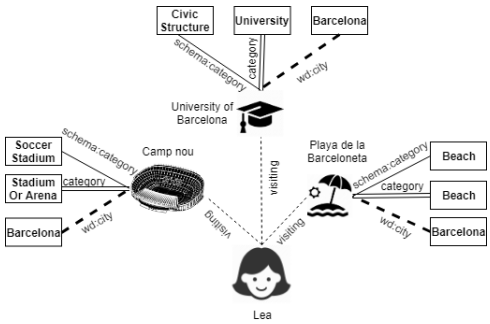


Figure 3: Semantic Trails Knowledge Graph

Table 5: Semantic Trails Datasets

Feature Name	Type	Range or #different values
Trail ID	Numerical	{1..Number of Trails}
User ID	Numerical	{1..Number of Users}
Venue ID	Categorical	~ 4.4 Million
Venue Category	Categorical	934
Venue Schema	Categorical	162
Venue City	Categorical	43833
Venue Country	Categorical	207
Time Stamp	Date	2012-04-03 To 2018-10-19

4.3 Deep Component

As shown in Figure 1, the deep component is a feed-forward neural network, that takes as input the one-hot encoded vector of the traveler and the city (t,c) and transform these two vectors into low-dimensional and dense vectors through the embedding layer which is a single layer perceptron, whose weights are initialized randomly. Weights are updated in backpropagation phase. The user demographics information plus the city knowledge graph and textual embeddings are concatenated with the traveler and the city embeddings to form the input of the deep component.

4.4 Factorization machines Component

In [7], the authors modeled the first order feature interaction factorization machines term by using two layers. The first layer takes as input the vector $\mathbf{x}$ corresponding to contextual information, and create an embedding vector of each feature of $\mathbf{x}$. More formally, the first layer computes a vector $\mathbf{v}_i \in \mathbb{R}^k$ for each feature i, where k is the dimension of features vector. In the second layer, the following pooling operation is performed $\mathbf{x}$:

$$f(x) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} x_i \mathbf{v}_i \odot x_j \mathbf{v}_j \tag{3}$$

where, $\odot$ denotes the element-wise product.

As shown in Figure 1, we use the same two layers in order to compute the factorization machines feature vectors interaction term.

4.5 Deep Knowledge Factorization machines

The obtained vectors from the deep component and the factorization machines component are concatenated and fed in a multilayer perceptron that contains different hidden layers. In each hidden layer l, we perform the following computation:

$$a[l] = Relu(\mathbf{W}[l-1]^T \mathbf{a}[l-1] + b[l-1]) \tag{4}$$

where, $Relu(x) = max(0, x)$ is the rectified linear unit function. It is used as the activation function for each layer of the multilayer perceptron. While there are other functions that can be used as activation function (sigmoid or hyperbolic tangent), Relu function is proven to avoid vanishing gradient problem, and showed better results in the experiments. $a[l-1], W[l-1], b[l-1]$ are respectively the activations, weights and bias of the previous layer (layer l-1).

Finally, at the end of the last hidden layer L, we compute the prediction $\hat{y}_{tc}$ by applying a sigmoid function to restrict the value between 0 and 1 which represents the probability to recommend the city c to the traveler t:

$$P(t, c|\mathbf{X}) = \hat{y}_{tc} = \sigma(\mathbf{h}^T \mathbf{a}[L]) \tag{5}$$

where, $\sigma(x) = \frac{1}{1+e^{-x}}$, h is the weight vector of the last neuron, $\mathbf{X}$ is the input vector of the multilayer perceptron.

The objective function of the backpropagation algorithm is to minimize the logistic loss defined as the negative log-likelihood of the observation (1 if traveler t went to city c, 0 otherwise) given the model's predictions:

$$\begin{aligned} Loss(\hat{y}_{tc}, y_{tc}) &= -log(P(\hat{y}_{tc}|\mathbf{W}, b, \mathbf{h})) \\ &= -(\sum_{u=1}^{n} \sum_{i=1}^{m} y_{tc} \times (log(\hat{y}_{tc}) + (1 - y_{tc}) \times log(1 - \hat{y}_{tc})) \end{aligned} \tag{6}$$

where, $P(\hat{y}_{tc}|\mathbf{W}, b, \mathbf{h})$ is the likelihood function of $\hat{y}_{tc}$, n is the number of users, and m the number of items.

5 EXPERIMENTS

In this section, we will try to give answers to the following research questions with empirical experiments:

(i) **RQ1**: What is the contribution of the deep component?
(ii) **RQ2**: What is the contribution for each input used in the deep component: traveler demographics data, city embeddings?
(iii) **RQ3**: What is the contribution of factorization machines component?
(iv) **RQ4**: How our model perform in comparison with baseline models?
(v) **RQ5**: How the performance of our model is affected by the hyper-parameters?

5.1 Experimental Setup

In this subsection, we present the different settings used to conduct the experiments as well as the baseline models used to perform the empirical comparison.

Dataset: We evaluate our model with the dataset obtained in the section 3.2. The characteristics of the dataset are shown in Tables 3 & 4.

Training & Test Sets: The recommendation task consists in predicting the next trip for a given traveler based on his previous trips. Hence, the dataset must be split in such a way that the test set must contain the last trip for each traveler. To do so, we adopt the leave-one-out strategy used in [8]. Formally, for each traveler, his last trip is used in the test set and the remaining trips are kept for the training set. Finally, n_s random cities where the traveler never went to are considered as negative samples. The experiments showed that n_s was performing well for a value of 3.

Evaluation Metrics: The output of the recommender system is a ranked list that contains all the cities where the user never went to in addition to the city in the test set. Hence, a good recommender system will rank this city in the test set at the top-K (we set K=10). To evaluate, our recommender system, we used two metrics defined as follows:

- **HR@K**: Hit Ratio metric measures whether the relevant city is within the Top K in the ranked list returned by the predictive model:

$$HR@K = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{K} hit(t, c_j) \tag{7}$$

- **MRR@K**: Mean Reciprocal Rank metric is used to measure how well the predictive model ranked the relevant city against the irrelevant ones:

$$MRR@K = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{K} \frac{1}{rank(t, c_j)} \tag{8}$$

where, $hit(t, c_j$ is equal to 1 if the traveler t visited the relevant city c_j, $rank(t, c_j)$ is the position of the relevant city c_j in the ranked list, n is the number of travelers.

Baseline Models: We compare our model Deep Knowledge Factorization Machines with a set of baseline methods that include collaborative filtering methods, factorization machines model and also two state-of-the-art deep learning based recommender systems. All the baselines are summarized below:

- **MostPop:** Cities are ranked by their popularity. The popularity of a city is calculated by the number of visits by all the travelers. The top-k popular cities are proposed to every traveler.
- **ItemKNN [23]:** This is a neighborhood based collaborative filtering algorithm based on items. The idea is to compute an item-item similarity matrix based on Pearson correlation coefficient.
- **ImplicitMF [9]:** This method was proposed to deal with implicit feedback data when using Matrix Factorization algorithm. They added a weight term to consider the confidence of an item and proposed an alternating least square algorithm to learn user's and item's latent vectors.
- **BPRMF [21]:** This method tailored for implicit feedback is a Matrix Factorization based method, but rather than minimizing the predicted "rating" as done in classical MF, it minimizes the pairwise ranking loss.
- **FM [20]:** This method was proposed by Rendle in order to incorporate contextual data in the recommendation. Instead of computing only users' and items' latent vectors, it computes also features latent vectors.
- **WDL [2]:** Wide & Deep Learning model combined a deep component (similar to the deep component used in our model) plus a wide component which can be seen as a linear model that computes cross products between input features. For this baseline, we used only the deep component. Indeed, The wide component is not adapted for our case as it requires the impression items.
- **NCF [8]:** Neural Collaborative Filtering is a state-of-the-art collaborative filtering approach. It combines the (user, item) interaction as input of a multilayer perceptron and a single layer perceptron that models the matrix factorization method.
- **NFM [7]:** Neural Factorization machines is a state-of-the-art model for context-aware recommendation. The factorization component used in our model represents a part of the neural factorization machines, the other part is a multilayer perceptron to which we add the linear term of factorization machines formula.

Implementation Framework & Parameter Settings: Our model plus all the baselines were implemented using Python and Tensorflow library[9]. The hyper-parameters of all the models were tuned using grid-search algorithm. First, we initialized all the weights randomly with a Gaussian Distribution ($\mu = 0$, $\sigma = 0.01$), and we used mini-batch Adam optimizer [11]. It is worth mentioning that other optimizers could be used in order to minimize the loss function defined in (6), however, Adam Optimizer has shown to be the most efficient in time and also accuracy. We evaluated our model using different values for hyper-parameters: the size of traveler and city embedding layers: $E_size \in \{32, 64, 128\}$, the features vector size of the factorization machines component: $k \in \{8, 16, 32, 64\}$, the batch size: $B_size \in \{64, 128, 256, 512, 1024\}$, the number of epochs: $epochs \in \{5, 10, 15, 20\}$ and the learning rate: $l_r \in \{0.001, 0.005, 0.006, 0.008, 0.1\}$.

5.2 Results and Discussion

Deep component performance: In the table 6, we present the results of the deep component using the metrics we have defined in the previous subsection. We can notice that the traveler demographics information (DCU) remarkably improved the performance of deep component that has only the traveler and city embedding as input (DC). The scores of HR@10 and MRR@10 increased by 15%. As for the city embeddings, we can notice that using the textual embeddings (DCTEI) improved the results by 9% and 6%. When considering the traveler demographics in addition to the textual city embedding (DCUTEI) the results improved by 27% for HR@10 and 28% for MRR@10. Finally, when concatenating all the deep component input, we improved our scores by 30% and 28%. The batch size used is 256, the number of epochs used is 8. Even if the loss defined in equation 6 decreased after 5 epochs for both the training and validation set, the neural network is over-fitting and the metrics HR@10 and MRR@10 decreased. Finally we used 0.006 as learning rate and 128 as the size of the traveler and city embedding layers. It is worth to notice that the hidden layer size

[9]Python Tensorflow API: https://www.tensorflow.org

Table 6: Contribution of each input in the deep component

Model	HR@10	MRR@10	#Layers	1st Layer size
DC	0.66	0.32	2	256
DCTEI	0.72	0.34	2	512
DCKGEI	0.79	0.36	2	512
DCU	0.76	0.37	2	512
DCUTEI	0.84	0.41	2	1024
DCUKGEI	0.84	0.40	2	1024
DCUI	**0.86**	**0.41**	3	1024

Table 7: Contribution of each input in the deep knowledge factorization machines

Model	HR@10	MRR@10	#Layers	1st Layer size
DKFM_CTXT	0.72	0.34	2	256
DKFMTEI	0.79	0.37	2	512
DKFMKGEI	0.80	0.38	2	512
DKFMU	0.82	0.38	2	512
DKFMUTEI	0.84	0.41	2	1024
DKFMUKGEI	0.85	0.42	2	1024
DKFM	**0.88**	**0.44**	3	1024

and the number of layers used vary over the input size as shown in the table 6.

Factorization machines contribution: The deep knowledge factorization machines was implemented by combining the feature vectors obtained from the factorization component by using the number of passengers in the booking in addition to the departure day of week as contextual data, and the deep component. Similarly to the table 6, we report the contribution of each input in the table 7. It is worth to notice that adding the contextual data increase the score of HR@10 and MRR@10 (DKFM_CTXT).

HR@10 and MRR@10 increased by 10% and 9% respectively when adding the textual embeddings (DKFMTEI). As for the knowledge graph embeddings, both scores increased by 11% (DKFMGEI). When considering the user demographics data (DKFMU), one can notice that the results improved by 13% and 11%. Finally, when considering all the input of our model DKFM, the results were 22% for HR@10 and 30% for MRR@10 better than the DKFM with only contextual data. We tested different values of k from 8 to 128, and we did not notice any change neither on the test loss, nor on the metrics HR@10 and MRR@10.

DKFM against baseline Models: We have measured HR@10 and MRR@10, for the different baseline models implemented and for our model DKFM. We present the results in Figure 4. As shown in Figure 4, our model outperforms the collaborative filtering methods demonstrating the importance of adding the city embeddings, traveler demographics data and the contextual data. It also shows a slight improvement over wide and deep learning and factorization machines where one is using city embeddings and traveler demographics data and the other is using contextual data. Considering that training time is also an important aspect to consider when doing recommendation, we measured training times for both DKFM and WDL models. For each epoch, the training time is equal to **24** seconds for DKFM and **15** seconds for WDL. For our experiments, we used an NVIDIA Tesla K40C GPU with 12 GB of memory. Finally we compute the metrics HR@K and MRR@K for different values of K and we report the results in figure 5.

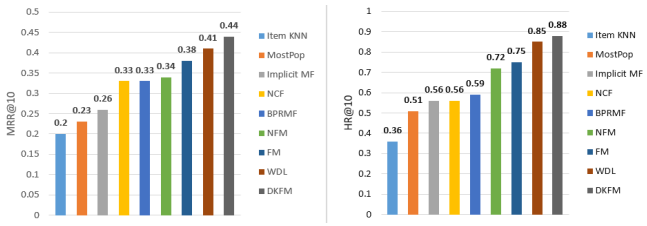


Figure 4: HR@10 & MRR@10

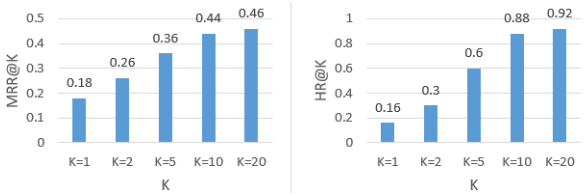


Figure 5: HR@K & MRR@K for DKFM

6 CONCLUSIONS AND FUTURE WORK

In this work, we have presented a neural network based approach to predict next trip destination for travelers. We have leveraged two external resources in order to enrich the characteristics of our destinations represented by a city. Our recommender system could be used for offline recommendations by sending emails to travellers, as well as for online recommendations on the airline's website to inspire them. We publish our code as open source in order to ease reproducibility: https://gitlab.eurecom.fr/amadeus/DKFM-recommendation. We conducted several experiments to address our research questions:

RQ1: What is the contribution of the deep component? The results of the experiments presented in table 6 show that when using Deep component with traveler demographics and city embeddings enhance significantly the next trip recommendation. Indeed, as shown in Figure 4, WDL outperforms all the collaborative filtering methods and also factorization machines model.

RQ2: What is the contribution for each input used in the deep component: traveler demographics data, city embeddings? The experiments demonstrated that the traveler demographics data improves remarkably the performance of the deep component. The textual embedding also improves the performanc of the deep component, but less than the traveler demographics. Finally, it showed that the knowledge graph embeddings improve more the score of the metrics in comparison with the two other inputs. Feeding DKFM with more relevant features improve the results by allowing the Deep Neural Network to learn more associations by combining the input features. However our model is considered as a black box

Table 8: Tuned Hyper-parameter for DKFM

Hyper-parameter	E_size	l_r	k	B_size	$epochs$
Value	128	0.06	8	256	8

in the sense that we are not able to say which input features were most important. This is in contrast with a regression or decision tree model where the information gain gives the information about the important features.

RQ3: What is the contribution of factorization machines component? Factorization machines component showed better results than collaborative filtering methods. However, the classical factorization machines got better results. This can be explained by the linear term that we do not use in our model. Indeed this term models the strength of each features.

RQ4: How our model perform in comparison with the baseline models defined later in this section? Our model outperforms all the collaborative filtering methods plus the deep learning methods [2, 7]. One can notice that the baseline MostPop has a good score for HR@10. Indeed, as there are only 119 cities to recommend (a few number of items), recommending the 10 most popular cities is at least one time out of two relevant.

RQ5: How the performance of our model is affected by the hyper-parameters? We ran grid-search on all the DKFM's hyper-parameters. The tuned hyper-parameter for our dataset are presented in table 8. We compared range of values for the size of the embedding layers, where 128 showed to be the value that has the highest HR@10, and we also compared different values for k: the size of feature vector from factorization machines component. For all the values we compared (see Section 5.2), HR@10 and MRR@10 did not vary.

In the future, we will first explore new data sources such as images that would help to enrich cities' characteristics. Next, we will investigate how we could improve the performance of our DKFM model by exploring new loss functions such as pairwise loss used in BPRMF [21]. We will also experiment the use of similarity measures inside the neural network such as cosine similarity. Finally, we plan to evaluate the performance of our recommender system for business travels.

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *26th International Conference on Neural Information Processing Systems - Volume 2*. Curran Associates Inc., USA, 2787–2795.

[2] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *1st Workshop on Deep Learning for Recommender Systems*. ACM, New York, NY, USA, 7–10.

[3] Thierry Delahaye, Rodrigo Acuna-Agost, Nicolas Bondoux, Anh-Quan Nguyen, and Mourad Boudia. 2017. Data-driven models for itinerary preferences of air travelers and application for dynamic pricing optimization. *Journal of Revenue and Pricing Management* 16, 6 (01 Dec 2017), 621–639. https://doi.org/10.1057/s41272-017-0095-z

[4] Gintare Karolina Dziugaite and Daniel M. Roy. 2015. Neural Network Matrix Factorization. *CoRR* abs/1511.06443 (2015).

[5] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In *22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 855–864.

[6] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 1725–1731.

[7] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 355–364.

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 173–182.

[9] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 263–272.

[10] Houda Khrouf and Raphaël Troncy. 2013. Hybrid Event Recommendation Using Linked Data and User Diversity. In *7th ACM Conference on Recommender Systems*. ACM, Hong Kong, China, 185–192.

[11] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[12] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. In *NIPS*.

[13] Julia Kiseleva, Melanie J.I. Mueller, Lucas Bernardi, Chad Davis, Ivan Kovacek, Mats Stafseng Einarsen, Jaap Kamps, Alexander Tuzhilin, and Djoerd Hiemstra. 2015. Where to Go on Your Next Trip?: Optimizing Travel Destinations Based on User Preferences. In *38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 1097–1100.

[14] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJvJXZb0W

[15] Augusto Q. Macedo, Leandro B. Marinho, and Rodrygo L.T. Santos. 2015. Context-Aware Event Recommendation in Event-based Social Networks. In *9th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 123–130.

[16] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *International Conference on Language Resources and Evaluation (LREC 2018)*.

[17] Diego Monti, Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, and Maurizio Morisio. 2018. Semantic Trails of City Explorations: How Do We Live a City. *CoRR* abs/1812.04367 (2018).

[18] Enrico Palumbo, Giuseppe Rizzo, and Raphaël Troncy. 2017. Entity2Rec: Learning User-Item Relatedness from Knowledge Graphs for Top-N Item Recommendation. In *Eleventh ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 32–36.

[19] Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, Elena Baralis, Michele Osella, and Enrico Ferro. 2018. Translational models for item recommendation. In *ESWC 2018, 15th European Semantic Web Conference*. Heraklion, GREECE.

[20] Steffen Rendle. 2010. Factorization Machines. In *2010 IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 995–1000.

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Arlington, Virginia, United States, 452–461.

[22] Francesco Ricci. 2002. Travel recommender systems. *IEEE Intelligent Systems* 17, 6 (2002), 55–57.

[23] Badrul Sarwar, George Karypis, Joseph A Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *10th International Conference on World Wide Web, WWW 2001*. ACM, Inc, New York, NY, USA, 285–295.

[24] Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Long-Kai Huang, and Chi Xu. 2018. Recurrent Knowledge Graph Embedding for Effective Recommendation. In *12th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 297–305.

[25] Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.

[26] Wolfgang Wörndl, Alexander Hefele, and Daniel Herzog. 2017. Recommending a sequence of interesting places for tourist trips. *Information Technology & Tourism* 17, 1 (2017), 31–54.

[27] Carl Yang, Lanxiao Bai, Chao Zhang, Quan Yuan, and Jiawei Han. 2017. Bridging Collaborative Filtering and Semi-Supervised Learning: A Neural Approach for POI Recommendation. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 1245–1254.

[28] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 353–362.

GViz - An Interactive WebApp to Support GeoSPARQL over Integrated Building Information

Kris McGlinn
kris.mcglinn@adaptcentre.ie
ADAPT Centre, Trinity College
Dublin, Ireland

Darragh Blake
darragh.blake@adaptcentre.ie
ADAPT Centre, Trinity College
Dublin, Ireland

Declan O'Sullivan
declan.osullivan@adaptcentre.ie
ADAPT Centre, Trinity College
Dublin, Ireland

ABSTRACT

Linked data (LD) is a technology to support publishing structured data on the web so that it may be interlinked. Building Information Modelling (BIM) is a key enabler to support integration of building data within the buildings life cycle (BLC). LD can therefore provide better access and more semantically useful querying of BIM data. The integration of BIM into the geospatial domain provides much needed contextual information about the building and its surroundings, and can support geospatial querying over BIM data. Creating GeoSPARQL queries for users who are non experts in semantic web technologies can be a challenge. In this paper we present a visualization tool built upon HTML5 and WebGL technologies that supports queries over linked data without the need to understand the resulting SPARQL queries. The interactive web interface can be quickly extended to support new use cases, for example, related to 3D geometries. The paper discusses the underlying data management, the methodology for uplifting several open data sources into Resource Description Framework (RDF), and the front-end implementation tested over a sample use case. Finally some discussion and future work is given, with a focus on how this tool can potentially support BIM integration.

CCS CONCEPTS

• **Information systems** → **Resource Description Framework (RDF)**; **Geographic information systems**; *RESTful web services*; • **Theory of computation** → **Data integration**; • **Human-centered computing** → Usability testing.

KEYWORDS

Linked Data, Building Information Modelling, GeoSPARQL, WebApp, Usability

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316536

ACM Reference Format:
Kris McGlinn, Darragh Blake, and Declan O'Sullivan. 2019. GViz - An Interactive WebApp to Support GeoSPARQL over Integrated Building Information. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316536

1 INTRODUCTION

Access to reliable structured data plays a central role in supporting existing and future services for managing smart and sustainable buildings and cities [7]. Building Information Modelling (BIM) has been identified as a key enabler to support integration of building data across its life-cycle [8] and with other data sources, such as those related to the buildings geolocation and external environment [29]. The combination of BIM and Linked Data (LD) [16] has the potential to meet the requirements for storing, sharing and interlinking BIM with other data sources, that is, once the data is represented or tagged using Resource Description Framework (RDF) and then published on the web.

Due to a lack of access to detailed BIM models used during the design and construction of buildings, e.g. models stored in the BIM standard Industry Foundation Classes, existing research is looking at methods to iteratively build BIM models using available open data, for example, the authoritative data made available by the Ordnance Survey Ireland (OSi, Ireland's national mapping agency) [24]. OSi data includes over 3.5 million buildings, consisting of GIS data (polygon foot print, geodetic coordinate), and also additional building data such as form, function and status. This data meets some basic requirements for BIM. Combined with other open data such as that provided by the Irish government on data.gov.ie, already a rich set of data sources are available for integration.

The challenge remains how to demonstrate the benefits of this LD approach. Several different clients are available for querying RDF triplestores [31], some having a visual component [30]. These types of tools that support users to do queries over several data sources can provide a gateway into LD based integrated BIM, and visualization can often make the results easier to digest for users not familiar with SPARQL. In this paper we present a WebApp for searching over multiple LD data sets describing different building

data sources taken from data.gov.ie and the OSi's data. The WebApp makes use of GeoSPARQL, an Open Geospatial Consortium (OGC) standard that defines a vocabulary for representing geospatial data on the Semantic Web, as well as specifying an extension to the SPARQL query language for processing that geospatial data. Using this underlying technology to manage the data, the WebApp makes use of HTML5, WebGL and Google Maps to provide a means to run polygon based geospatial queries over a map of Ireland, selecting distinct areas and querying buildings in those areas for available data, presented back to the user as a map overlay.

The data used is uplifted into RDF using a tried and tested approach based on R2RML [35] applied to both the data.gov.ie data sets (converting from the available CSV) and also the conversion of the OSi Prime2 data base [14]. The paper presents a web based usability evaluation of the tool that explores users doing several tasks using the key feature of the tools, the geospatial querying. The paper concludes with a discussion of the results of this evaluation, and possible extensions of the tool and methods to support further integration of BIM through the use of geolocation.

2 STATE OF THE ART

Building Information Modelling

Building Information Modelling (BIM) describes an integrated data model for storing all information relevant to the buildings life cycle. This can include a 3D model of an architectural design, electrical installations, fire protection, occupancy, energy consumption, costs, etc [8]. BIM goes further than just providing consistent representation of objects; it also defines object parameters and relations to other objects. Open BIM has the potential to support new and innovative services to support intelligent automation, navigation, energy efficiency and sustainability, but faces several challenges, related to; standardization, data interdependence, data access, security [1] and also the willingness of BIM developers and owners to share their BIM data [24].

For the true potential of BIM to be realized, it is important that developers of new services are given access to available, open and authoritative BIM. Due to the limited availability of BIM data in traditional formats, such as Industry Foundation Classes [24], GIS and other data sources can be used to construct rudimentary BIM models based on a buildings geolocation, and other attributes such as: address, form, function and other openly available properties. Nonetheless, these open data sources remain scattered between different services including the aforementioned data.gov.ie, as well as DBpedia [12]. Linking these data sources with the authoritative OSi dataset, provides an important step toward making BIM available to a wider range of users [24].

Linked Data, BIM and Graphical Information Systems

Linked Data (LD) is an approach to expose, share, and connect related data, not previously linked, on the Web [16]. In modern Architecture, Engineering and Construction, data related to different domains such as building geometry and topology data, sensor data, behaviour data, are generated and consumed across BLC stages. The combination of BIM and LD has the potential to meet the requirements for storing and sharing this data. However, the data must be represented as or at least tagged using RDF. GIS systems are information systems with an additional geo-reference [20], a requirement for spatial analysis. Spatial information includes coordinates and relationships between features, as well as additional non-spatial attributes. GeoSPARQL is an Open Geospatial Consortium (OGC) standard that defines a vocabulary for representing geospatial data on the Semantic Web, but also specifies an extension to the SPARQL query language for processing that geospatial data. Research efforts around GeoSPARQL are still ongoing, for example, to explore the conversion of non-RDF data into GeoSPARQL [4], or support for enabling on-the-fly GeoSPARQL-relational database translation [9]. As discussed previously, the OSi's authoritative geospatial data for boundaries has been converted into GeoSPARQL and made available through the GeoHive (data.geohive.ie) domain.

Linking BIM with geospatial data is also an active area of research. In [18] the integration of the BIM standard Industry Foundation Classes (IFC) is explored to support site planning, in terms of localization of materials and services appropriate for optimized productivity of a construction project. IFC is a complex schema originally developed for use by the Architecture, Engineering and Constructions domain for buildings, and making it more accessible for querying can potentially open up new markets and novel tools [24]. Other research has investigated the conversion of standards like IFC directly into CityGML, a standard for describing data related to cities and districts and includes capabilities for describing buildings also [15].

LD-based GIS Visualisation Tools

Several triplestores have been developed to support GeoSPARQL, for example: Parliament, Strabon, Stardog and Fuseki [2, 19, 32, 34]. Each of these have web clients used to access and query the RDF data. Other clients are developed specifically to support editing [11], [10] or querying and visualizing results of triplestores [17, 33, 37]. Collectively, these tools provide a wide range of different functions related to autocomplete, multiple endpoints, query retention, file upload, results download etc. [31].

A popular tool for testing geospatial queries is Yasgui, which has seen features it supports being integrated with a wide range of other related tools [31]. One interesting aspect and a potential contributor to its popularity may be its additional rendering capabilities for visualization of geosparql geometries on a world map [30], allowing users to quickly see the results of their geospatial queries. What none of these tools offer is the creation of geospatial queries based on interaction with the map and the results. Also, the integration of 3D geometries has not yet been considered, and the potential to extend into 3D, so that both 3D and 2D can be queried side by side, could provide new capabilities and support for a wider range of use cases.

The remainder of this paper presents the GViz WebApp that supports GeoSPARQL and the visualization of well-known text (WKT) geospatial geometries, configuration of geospatial queries through an interactive interface, and is built on extensible technologies such as HTML5, JS and WebGL. These features together mean GViz has the potential to support a wide range of existing and future use cases for querying and displaying geospatial geometries.

3 DESIGN

The GViz tool has three main requirements, that it:

(1) supports platform independent configurable geospatial functions to query and visualise integrated data with a geospatial component on the web
(2) be flexible and extensible so as to support existing and future requirements
(3) be usable: i.e. users can efficiently and effectively achieve the tasks it supports whilst also finding the tool to be satisfying to use

To meet 1) the tool design must enable the integration of an RDF triplestore with GeoSPARQL support. The web client itself must enable a user to select a region on a map, and run a query over that region returning any geospatial results within. Additionally it must support the user to access and display data through the interactive map, based on properties of the data, for example, in the case of a building to display data properties about the building. The triplestore must therefore support GeoSPARQL functions. To meet 2) the tool is designed to make use of modern web technologies, such as HTML5 and WebGL, that can support existing and future requirements through highly modular and configurable JS libraries. To meet 3) the tool must be developed with usability in mind, making the completion of actions intuitive and user friendly. Towards that end, this paper presents an initial usability evaluation of the tool set in section 6.

Data Uplift Methodology

As the GViz tool requires data to be represented as RDF, this section briefly discusses the process currently employed to convert structured non-RDF data into RDF. This process, called uplift [14] has been successfully employed in the conversion of several different data sets most notably the conversion of the Ordnance Survey Ireland (OSi) Prime2 data base boundary data sets, but also OSi data related to Ireland's buildings alongside DBpedia data [24], Irish Central Statistics Office data, and BIM data extracted from IFC [23]. More recently OSi building data sets (containing over 3.5 million building representations) are being converted. To date about 80 thousand buildings are now represented as RDF, with the remainder due to be converted within the next 3-6 months. The properties being converted into RDF include the buildings name (Irish and English), geolocation, and form and function. In future releases these will include more information (e.g. addresses, status, etc.), including provenance information, i.e. historical values and information about who and when changes were made.

The uplift methodology [14] makes use of the R2RML mapping language, and supports conversion of any tabular data source (e.g. CSV, SQL, etc.). The process of conversion consists of 4 main steps. The first an analysis to understand the data source(s), to define mappings to existing vocabulary, and where no mapping exist, to propose extension to these or develop new ontologies. Second, the development of the R2RML mappings to convert the data to RDF and the running of the R2RML process to do this conversion. And finally, the deployment of the data on a triplestore with an exposed endpoint for querying of the data. In step three, testing of the data is also conducted to ensure that the R2RML mappings have generated the data correctly. This includes testing through known tools, such as Yasgui or GViz (for quick visualize confirmation), or more in depth analysis using tools such as luzzu [13].

4 IMPLEMENTATION

The GViz tool has two main components; the back end triplestore and the front end web client. There is also a component to manage secure authentication, discussed here only briefly.

Back-end

To support the type of polygon selections that the GViz tool is to enable, the underling triplestore must support GeoSPARQLs sf:Within. Stardog, Strabon and Parliament all support this functionality. Fuseki currently does not support sf:Within, and is also quite poorly documented making it difficult to set up [3], although it is believed an update to the Fuseki server will mean that future versions will be both easier to deploy, and also support the full range of

GeoSPARQL's functionality. Stardog requires a commercial licence to use, and Strabon, like Fuseki, has a complex set of instructions to set up with poor documentation. Parliament is by far the easiest of the non-commercial GeoSPARQL supporting triplestores to run and enable indexing of geospatial data. This can be done through its web-client with relative ease, making Parliament a good test triplestore. Parliament appears to have ceased active development as of 2016, and so, if the extension to Fuseki materialize, Fuseki may well become a better option. Once you run the triplestore, data is accessed through its endpoint.

Front-end

The GViz front-end supports querying of triplestores through hardcoded endpoints in JS, although this can be easily extended to support user configurable endpoint through an interface, similar to that supported by Yasgui. Access by the client to the underlying triplestore is handled using AJAX. Calls to the triplestore are managed through interaction with the front-end web client, running on the client browser, and built upon JS, HTML5, Google Maps API and WebGL.

Google Maps API was chosen as it provides access to googles maps, as well as providing additional functionality, such as capabilities to add markers to geospatial points, and a drawing manager to support drawing polygons (used for selection). WebGL supports drawing points and 2D polygons, as well as polygon meshes if required. HTML5 and JS further support select boxes and pop up information windows, for example, to filter data or to provide feedback from point and click selections of results. Fig. 1 shows the front-end as an area is selected (black lines), and the resulting geospatial coordinates returned by the select and query (red dots). In the next section we describe a test case for the tool in which four data sources were chosen, uplifted, and then made available for query using GViz. It should be noted that the current implementation of GViz is being run on a web server using node.js. This was to enable authentication, so that different data sets can be made available depending on the log-in. This is implemented using express and mongodb. No support for https can be given though, as Parliament does not support https.

5 GVIZ TEST CASE

To validate the approach four data sets were chosen for integration with the visualization app. The data sets adhere to two criteria; 1) are available in a structured format (e.g. CSV, XML) and 2) have an explicit spatial component. Three data sets from data.gov.ie were chosen that relate to building data. These are:

- https://data.gov.ie/dataset/churches-and-convents1
  - Properties: ID, Name, Address, Organisation, GeoSpatial Point (Longitude and Latitude)
- https://data.gov.ie/dataset/schools1
  - Properties: ID, Address, Phone, School Level, Mixed Status, Fee Paying, GeoSpatial Point
- https://data.gov.ie/dataset/roscommon-thatch-building-survey
  - Properties: ID, Address, Image URI, GeoSpatial Point

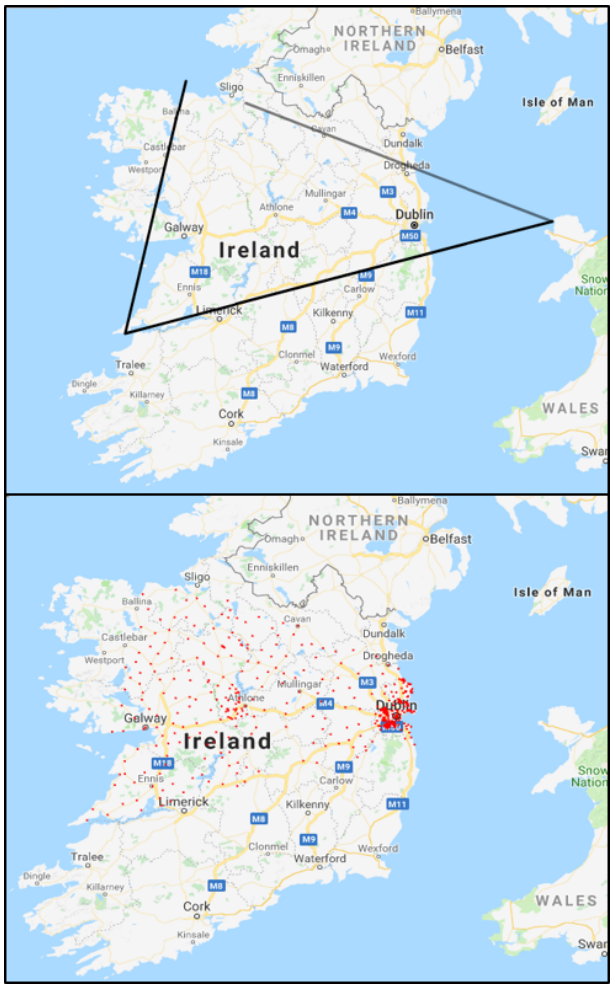


Figure 1: GViz - Polygon area based selection

An additional data set was also used, provided directly by OSi (all Garda station buildings in Ireland) and includes the OSi buildings URI and additional data for form and function, information that is maintained by OSi. The OSi data set is generated based on a comprehensive analysis of the data sources [24], resulting in an ontology [22]. For the data.gov.ie data sets, the analysis was superficial, and no mappings were explored to existing vocabulary.

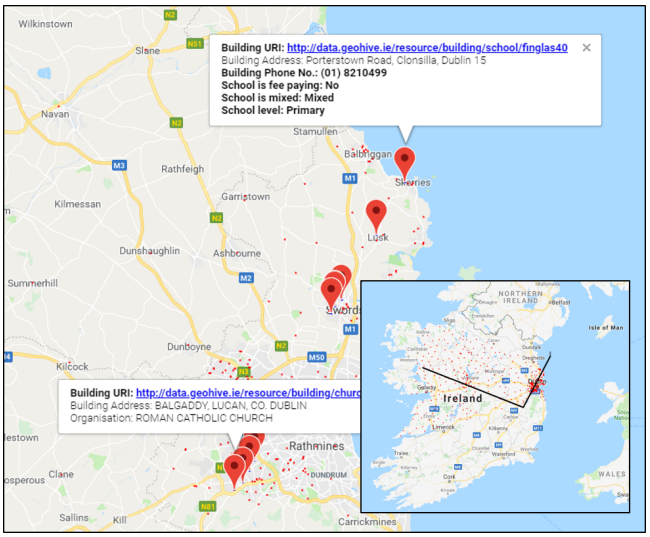


Figure 2: GViz - Polygon area based selection

Instead, a temporary namespace was created as a place holder for the properties, generated based upon the column names. Where columns across data sets shared similar semantics (e.g. Address), the same temp property was used, although in some cases the R2RML mappings need to concatenate columns into one property as Address could cover two or more columns, depending on the data set.

Listing 1: A sample SPARQL query to Parliament

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/
        function/geosparql/>
PREFIX osi: <http://ontologies.geohive.ie/osi#>
PREFIX geoff: <http://ontologies.geohive.ie/geoff#>
PREFIX temp: <http://ontologies.geohive.ie/temp#>

SELECT
    ?uri ?form ?function ?name ?address ?iuri
    ?phone ?org ?fee ?mixed ?level ?poly_geom
WHERE {
    ?uri geo:hasGeometry ?geom.
    OPTIONAL{?uri geoff:hasForm ?form.}
    OPTIONAL{?uri geoff:hasFunction> ?function.}
    OPTIONAL{?uri osi:hasName> ?name.}
    OPTIONAL{?uri temp:hasAddress> ?address.}
    OPTIONAL{?uri temp:hasImageURI> ?iuri.}
    OPTIONAL{?uri temp:hasPhone> ?phone.}
    OPTIONAL{?uri temp:hasOrganisation> ?org.}
    OPTIONAL{?uri temp:hasFeePaying> ?fee.}
    OPTIONAL{?uri temp:hasMixedStatus> ?mixed.}
    OPTIONAL{?uri temp:hasSchoolLevel> ?level.}
    ?geom geo:asWKT ?poly_geom
    .FILTER(geof:sfWithin(?poly_geom,
        "POLYGON(( -8.84875046875004 54.09632741172031
        , -9.62877976562504 52.66126550188105,
        -4.9486039843750405 53.32254658410386,
        -8.84875046875004 54.09632741172031))"
        ^^geo:wktLiteral )).
}
```

In Fig. 1 an area query is executed through the interface, returning all geospatial data within the area. Listing 1 shows the resulting SPARQL query generated by the select, and adapted for the RDF generated by the R2RML mappings. Fig. 2 shows a “select" being run after the initial query. The select functionality allows the user to draw a second line (or polygon) to intersect with points of interest. Each point intersected by the line displays a clickable marker. Upon clicking this further information on the building it represents is displayed, based upon the R2RML mappings. Some additional functionality is also possible using radio buttons, allowing further filtering based upon the form and function of the building. To see the full set of current functionality of the tool, there is an online video available here [21].

6 USABILITY EVALUATION

As usability is one of the key requirements of the GViz tool, we conducted a preliminary evaluation to determine usability of the key feature of the tool, i.e. the capability to do configurable GeoSPARQL queries over the data. The methodology for the evaluation has been applied in previous usability assessments [25–27]. Full evaluations consist of both a formative and summative evaluation. Summative evaluations are conducted when a tool is fully developed. As this is a prototype tool, here we present an initial formative evaluation. These are conducted during the development to mould or improve a product and therefore, results are expected to reflect the early development stage. Outputs of formative evaluations may include participant comments (attitude’s, sources of confusion, and reasons for actions) and other usability problems and suggested fixes determined through observation and feedback.

The main evaluation metric we evaluate is user satisfaction, measured by using the System Usability Scale (SUS). SUS is a simple ten-item scale giving a global view of subjective assessment of usability. The statements in SUS are chosen to identify extreme expressions or attitudes. SUS also provides a point structure to assign to the answers of a particular test rating overall satisfaction between 0 and 100. Bangor et al. [5] suggest that a score in the seventies should be deemed acceptable for a finished product, and those below still have usability issues of concern. With respect to the number of participants required to find all potential problems, this may vary according to the users, the tasks, and the system under test. At least a range between five and fifteen is required to evaluate sensitive parameters as depicted in [28, 36].

The evaluation is structured upon Common Industry Format (CIF). A CIF usability report must include; a description of the product/model, the goals of the test, the test participants, their background and the tasks they are to perform, the method for conducting the test, the experimental design

Questions	Strongly Agree (%)	Agree (%)	Neutral (%)	Disagree (%)	Strongly Disagree (%)	Total
I think that I would like to use this system frequently	0%	19%	44%	31%	6%	100%
I thought the system was easy to use	0%	38%	31%	31%	0%	100%
I found the various functions in this system were well integrated	0%	38%	38%	25%	0%	100%
I would imagine that most people would learn to use this system very quickly	13%	69%	6%	13%	0%	100%
I felt very confident using the system	6%	44%	25%	25%	0%	100%
Total	4%	41%	29%	25%	1%	100%
I found the system unnecessarily complex	6%	31%	25%	38%	0%	100%
I think that I would need the support of a technical person to be able to use this system	6%	6%	13%	38%	38%	100%
I thought there was too much inconsistency in this system	0%	6%	19%	75%	0%	100%
I found the system very cumbersome to use	0%	31%	31%	31%	6%	100%
I needed to learn a lot of things before I could get going with this system	0%	6%	19%	56%	19%	100%
Total	3%	16%	21%	48%	13%	100%

Figure 3: SUS Scores

of the test, the usability measures and the numeric results and analysis [6].

Goal, participants and backgrounds

The goal of this experiment is to assess the level of usability of the GViz tool for users with technical backgrounds when conducting a typical task related to the geospatial queries. 16 participants took part in this experiment. For an initial evaluation, this number was considered sufficient. These were taken from members of the ADAPT-TCD research centre and the OSi Ireland. A pre-questionnaire asked for some background on their roles in their organization. This broke down as follows, from ADAPT: 4 postdocs/research fellows, 4 Ph.D students, 2 project managers/coordinators, 1 research assistant, 1 academic and 1 software engineer. Of these, 9 are computer scientists, 2 project managers, and 2 linguists. From OSi 2 GIS IT professionals, 1 Industry representative all working in the domain of GIS.

Experimental description, tasks description, and technologies

The evaluation explored the participants completion of four tasks (below). These relate to the main feature of interest in the GViz interface:

(1) Select an area of the map to run a GeoSPARQL query to return buildings within
(2) Select returned buildings for further examination
(3) Click on the marker to see data about the building
(4) Filter the data

More on these tasks can be found here [21]. The technologies used have been presented in the implementation section.

Findings

The SUS scored 60 out of 100. Fig. 3 gives a breakdown of the SUS questionnaire scores as percentages. The participants were also asked 'What features of the tool, if any, did you like/dislike?', "What additional features would you like to see, if any?" and to provide any 'Further comments'. Positive comments about the features included "Once using the app, it is intuitive", "Display of red dots (geolocation) on map is clear", "Interesting way to select a region", "the ability to select regions" and two related to the "Filtering" of results. Negative comments included "The shown geo locations (red dots) from the area selection are too small if you zoom in", "Having to use a keyboard: right click is preferred if you're not a power user; having to draw polygons: good if you want to be very specific but most people aren't -> use a simpler selection. Marker dots are indiscernible to the visually challenged.", "Having to use a keyboard: right click is preferred if you're not a power user; having to draw polygons: good if you want to be very specific but most people aren't -> use a simpler selection. Marker dots are indiscernible to the visually challenged.", "Coloured markers - very hard to distinguish colours when you are colourblind", "The select tools are basic for a web mapping tool".

For extra features participants commented: "Button options as opposed to "s" or "d". Maybe a "drawing" option as opposed to clicking to draw a polygon? (free form as opposed to straight lines)", "Maybe some kind of "dragging" feature that could allow you to draw say a square around the city, rather than having to define the boundaries yourself." and "different markers (shape?) for different types", and "the ability to download the selection (URI,X,Y,form,Function) as CSV or similar". For further comments: "Second attempt when selecting all filters did not reveal any buildings", "Its hard to see what the exact application of the tool is. If it is to select features and read data from pop-ups from a users point of view it compares poorly with propriety (Arc) and open source alternatives (mapbox, leaflet). The fact that it is using linked open data may be of interest from a research point of view." and "This has potential". In the next section we address these findings and comments.

7 DISCUSSION, CONCLUSION AND FUTURE WORK

This work presented a prototype WebApp called GViz for providing configurable geospatial functions using GeoSPARQL

on an interactive map. The tool has been demonstrated using a test case of four sample open data sets taken from both the OSi building data and data.gov.ie data. A usability evaluation was also conducted to gather early feedback on the usability of the tool when selecting and querying this data. Overall satisfaction (SUS 60) is close to a 'C' passing grade, but still below what would be needed for a release version. Considering the early nature of this prototype, we nonetheless consider this a promising start from a usability perspective, and part of an iterative process of development and evaluation.

There were many comments on how the tool could be improved in future implementations. Some of these comments are related to implementation choices, for example, the display of geolocations as red dots or the interaction using keys on a keyboard, that can be easily addressed in newer versions. Other comments, such as those related to how geospatial queries are created, for example, using preset shapes (circles for example around a point) are design choices that have already been considered, and again can be implemented in future versions. Ideally, the generated queries should be available for display alongside the interactive map and polygon drawing, similar to the SPARQL query editing box in a tool such as Yasgui. Other comments from GIS professionals were related to alternative tools such as leflet and mapbox, with one not having a full understanding of the unique ability to select from multiple integrated data sets based on a single query. Nor that the tools given work with their own internal proprietary models and APIs, and as such, cannot be said to be open. Negative comments like these may have also influenced the final SUS score, and could be ameliorated with a better explanation in the instructions.

The tool has been shown to provide functionality to enable a user to select an area of interest and run geospatial queries over the data set. This goes beyond the functionality of existing SPARQL web clients, such as Yasgui, with respect to interactive queries and visualisation of data. The GViz tool can therefore potentially form part of a suite of tools to support non-expert users generating queries over geospatial data, providing visual feedback. This research opens to doors to several possible extensions of the tool, here we discuss in short some of the potential extensions to GViz and future research exploration.

Integration of BIM data based upon Geolocation

In previous work the potential for using geolocation as an identifier for buildings was explored [24]. This work showed that it is possible to determine, based upon some criteria (i.e. distance between points) that two buildings with a location that is very close can be considered to be the same building, thereby opening up the possibility of integrating the data from two sources. Future work will examine the GViz tool

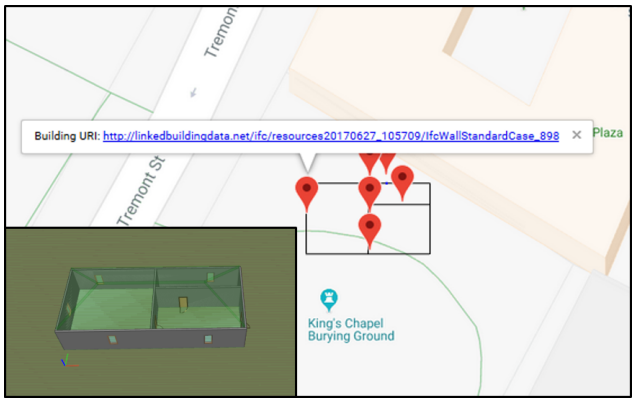


Figure 4: GViz - Visualisation of smallhouse.ifcowl walls after conversion to GeoSPARQL (selected using GViz)

as a potential for selecting an area, querying over several datasets, and then running an algorithm to determine if two points can be considered belonging to the same building. This will require consideration of the distances between buildings and the buildings geolocation, as it is possible if two buildings use a different method for generating its geolocation (for example, a centroid or the building entrance) the algorithm may determine that adjacent buildings are the same. Other properties will also be explored to improve the accuracy of the selection process, for example building name, address, building number, floor area etc.

Here GViz becomes a useful tool to do selects over areas, as the alternative would require all buildings across data sets to be compared, and in the case of OSis > 3.5 million buildings, such a process would require significant processing. In particular, it will give capabilities for users to align their models with authoritative OSi building URIs.

Integration of IFC models

In [23] it was shown that IFC can be uplifted and geospatial data extracted, thereby supporting geospatial functions over IFC data. This process can provide yet another data stream for integrate BIM and once again, the potential for building owners of BIM models to align and share aspects of their IFC models publicly, could potentially attract developers to develop new tools to support those owners managing and maintaining their buildings. This work is currently being extended to support the conversion of building elements, such as walls, into WKT, allowing a buildings floorplan to be converted in geospatial.

Fig. 4 shows an initial conversion of smallhouse.ifc to ifcowl with additional GeoSPARQL describing the buildings walls in 2D and displayed in GViz. The walls have then been selected, and the URI of the wall entity displayed. It should be

noted, smallhouse.ifc is a test model, and as such, its geospatial coordinates do not actually align with any existing physical site. We include it here because it is a popular test IFC model. The possibility therefore exists to move seamlessly from the geospatial domain into the building domain, aligning 3D BIM models with 2D geospatial data, and visualising this data through interactive WebApps like GViz.

ACKNOWLEDGMENTS

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (13/RC/2106) and is co-funded under the European Regional Development Fund. This work was partially funded by the Government of Ireland under the Open Data Engagement Fund.

REFERENCES

[1] Kereshmeh Afsaria, Charles M. Eastmann, and Dennis R. Sheldena. 2016. Cloud-based BIM Data Transmission : Current Status and Challenges.

[2] Apache. 2019. Fuseki: serving RDF data over HTTP. https://jena.apache.org/documentation/serving_data/. Accessed: 22-Jan-2019.

[3] Apache. 2019. Spatial searches with SPARQL. https://jena.apache.org/documentation/query/spatial-query.html. Accessed: 22-Jan-2019.

[4] Sören Auer, Jens Lehmann, and Sebastian Hellmann. 2009. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In *The Semantic Web - ISWC 2009*, Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 731–746.

[5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies* 4, 3 (May 2009), 114–123. http://dl.acm.org/citation.cfm?id=2835587.2835589

[6] Nigel Bevan. 1999. Common industry format usability tests. In *Proceedings of UPA'98, Usability Professionals Association.*

[7] Muhammad Bilal, Lukumon O. Oyedele, Junaid Qadir, Kamran Munir, Saheed O. Ajayi, Olugbenga O. Akinade, Hakeem A. Owolabi, Hafiz A. Alaka, and Maruf Pasha. 2016. Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics* 30, 3 (2016), 500 – 521. https://doi.org/10.1016/j.aei.2016.07.001

[8] Andre Borrmann, Markus Konig, Christian Koch, and Jakob Beetz. 2018. *Building Information Modeling: Technology Foundations and Industry Practice.* Springer International Publishing. ISBN: 978-3-319-92861-6.

[9] Diego Calvanese, Benjamin Cogrel, Sarah Komla-Ebri, Roman Kontchakov, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, and Guohui Xiao. 2017. Ontop: Answering SPARQL Queries over Relational Databases. *Semantic Web Journal* 8, 3 (2017), 471–487. https://doi.org/10.3233/SW-160217 Semantic Web Journal outstanding paper award for 2016.

[10] Stephane Campinas. 2014. Gosparqled - SPARQL Auto-Completion. https://github.com/scampi/gosparqled. Accessed: 22-Jan-2019.

[11] S. Campinas, T. E. Perry, D. Ceccarelli, R. Delbru, and G. Tummarello. 2012. Introducing RDF Graph Summary with Application to Assisted SPARQL Formulation. In *2012 23rd International Workshop on Database and Expert Systems Applications.* 261–266. https://doi.org/10.1109/DEXA.2012.38

[12] DBPedia. 2019. Structured content from Wikipedia project. http://wiki.dbpedia.org/. Accessed: 22-Jan-2019.

[13] Jeremy Debattista, Sören Auer, and Christoph Lange. 2016. Luzzu—A Methodology and Framework for Linked Data Quality Assessment. *J. Data and Information Quality* 8, 1, Article 4 (Oct. 2016), 32 pages. https://doi.org/10.1145/2992786

[14] Christophe Debruyne, Eamonn Clinton, Lorraine McNerney, Atul Nautiyal, and Declan O'Sullivan. 2016. Serving Ireland's Geospatial Information as Linked Data. In *Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference.*

[15] Yichuan Deng, Jack C.P. Cheng, and Chimay Anumba. 2016. Mapping between BIM and 3D GIS in different levels of detail using schema mediation and instance comparison. *Automation in Construction* 67 (2016), 1 – 21. https://doi.org/10.1016/j.autcon.2016.03.006

[16] Tom Heath and Christian Bizer. 2011. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology* 1, 1 (2011), 1–136. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

[17] F. Hogenboom, V. Milea, F. Frasincar, and U. Kaymak. 2010. *RDF-GL: A SPARQL-Based Graphical Query Language for RDF.* 87. https://doi.org/10.1007/978-1-84996-074-8_4

[18] Ebrahim P. Karan and Javier Irizarry. 2015. Extending BIM interoperability to preconstruction operations using geospatial analyses and semantic web services. *Automation in Construction* 53 (2015), 1 – 12. https://doi.org/10.1016/j.autcon.2015.02.012

[19] KRR&A. 2019. Strabon - Semantic Web Standards. http://www.strabon.di.uoa.gr/. Accessed: 22-Jan-2019.

[20] Xin Liu, Xiangyu Wang, Graeme Wright, Jack C. P. Cheng, Xiao Li, and Rui Liu. 2017. A State-of-the-Art Review on the Integration of Building Information Modeling (BIM) and Geographic Information System (GIS). *ISPRS Int. J. Geo-Information* 6 (2017), 53.

[21] Kris McGlinn. 2019. GVis Video Tutorial. https://www.scss.tcd.ie/~mcglink/video/tutorial/geovis/geovis.html. Accessed: 22-Jan-2019.

[22] Kris McGlinn. 2019. OSi Building Ontology. http://ontologies.geohive.ie/osi/building/. Accessed: 22-Jan-2019.

[23] Kris McGlinn, Christophe Debruyne, Lorraine McNerney, and Declan O'Sullivan. 2017. Integrating Building Information Models with Authoritative Irish Geospatial Information. In *ISWC 2017, the 16th International Semantic Web Conference.* 66–74. ISBN: 978-3-319-68204-4.

[24] Kris McGlinn, Christophe Debruyne, Lorraine McNerney, and Declan O'Sullivan. 2017. Integrating Ireland's Geospatial Information to Provide Authoritative Building Information Models. In *Proceedings of the 13th International Conference on Semantic Systems - Semantics2017*, Vol. 13. ACM Press, 57–64. https://doi.org/10.1145/3132218.3132223

[25] Kris McGlinn, Lucy Hederman, and Dave Lewis. 2014. SimCon: A context simulator for supporting evaluation of smart building applications when faced with uncertainty. *Pervasive and Mobile Computing* 12 (2014), 139 – 159. https://doi.org/10.1016/j.pmcj.2013.02.003

[26] Kris McGlinn, Kelvin Jones, Aitor Cochero, Anton Gerdelan, and Dave Lewis. 2015. Usability evaluation of an activity modelling tool for improving accuracy of predictive energy simulations. In *Proceedings of the 14th Conference of International Building Performance Simulation Association (IBSA).*

[27] Kris McGlinn, Baris Yuce, Hendro Wicaksono, Shaun Howell, and Yacine Rezgui. 2017. Usability evaluation of a web-based tool for supporting holistic building energy management. *Automation in Construction* 84 (2017), 154 – 165. https://doi.org/10.1016/j.autcon.2017.08.033

[28] Jakob Nielsen. 1994. Enhancing the Explanatory Power of Usability Heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94).* ACM, New York, NY, USA, 152–158. https://doi.org/10.1145/191666.191729

[29] Pieter Pauwels, Sijie Zhang, and Yong-Cheol Lee. 2017. Semantic web technologies in AEC industry: A literature overview. *Automation in Construction* 73 (2017), 145 – 165. https://doi.org/10.1016/j.autcon.2016.10.003

[30] Laurens Rietveld and Rinke Hoekstra. 2013. YASGUI: Not Just Another SPARQL Client. In *The Semantic Web: ESWC 2013 Satellite Events*, Philipp Cimiano, Miriam Fernández, Vanessa Lopez, Stefan Schlobach, and Johanna Völker (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 78–86.

[31] Laurens Rietveld and Rinke Hoekstra. 2015. The YASGUI Family of SPARQL Clients.

[32] SemWebCentral. 2019. Parliament High-Performance Triple Store. http://parliament.semwebcentral.org/. Accessed: 22-Jan-2019.

[33] Paul R. Smart, Alistair Russell, Dave Braines, Yannis Kalfoglou, Jie Bao, and Nigel R. Shadbolt. 2008. A Visual Approach to Semantic Query Design Using a Web-Based Graphical Query Designer. In *Knowledge Engineering: Practice and Patterns*, Aldo Gangemi and Jérôme Euzenat (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 275–291.

[34] Stardog. 2019. Stardog-The Knowledge Graph Platform for the Enterprise. https://www.stardog.com/. Accessed: 22-Jan-2019.

[35] W3C. 2019. R2RML: RDB to RDF Mapping Language. https://www.w3.org/TR/r2rml/. Accessed: 22-Jan-2019.

[36] Alan Woolrych and Informatics Centre. 2001. Why and when five test users aren't enough. In *Proceedings of IHM-HCI 2001 Conference*. 105–108.

[37] Martins Zviedris and Guntis Barzdins. 2011. ViziQuer: A Tool to Explore and Query SPARQL Endpoints. In *The Semanic Web: Research and Applications*, Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 441–445.

Geolocation in the Browser

From Google Gears to Geolocation Sensors

Thomas Steiner
Google LLC
20354 Hamburg, Germany
tomac@google.com

Anssi Kostiainen
Intel Corporation
02160 Espoo, Finland
anssi.kostiainen@intel.com

Marijn Kruisselbrink
Google LLC
San Francisco, CA 94105, USA
mek@google.com

ABSTRACT

Geolocation is arguably one of the most powerful capabilities of smartphones and a lot of attention has been paid to native applications that make use of it. The discontinued Google Gears plugin was one of the first approaches to access exact location data on the Web as well, apart from server-side coarse location lookups based on Internet Protocol (IP) addresses; and the plugin led directly to the now widely implemented Geolocation API. The World Wide Web Consortium (W3C) Geolocation API specification defines a standard for accessing location services in the browser via JavaScript. For a long time, developers have also demanded more advanced features like background geolocation tracking and geofencing. The W3C Geolocation and the Devices and Sensors Working Groups, as well as the Web Incubator Community Group (WICG), have addressed these demands with the no longer maintained Geofencing API specification for the former, and—with now (early 2019) resumed efforts—the in-flight Geolocation Sensors specification for the latter two groups. This paper first provides an in-depth overview of the historical development of geolocation in the browser and motivates privacy decisions that were made at the time, and then gives an outlook on current and future efforts, challenges, and use cases from both a technology as well as from a privacy angle.

CCS CONCEPTS

• **Information systems** → **Web applications**; • **Security and privacy** → **Browser security**.

KEYWORDS

Geolocation; Geofencing; Geotracking; Permissions; Privacy

ACM Reference Format:
Thomas Steiner, Anssi Kostiainen, and Marijn Kruisselbrink. 2019. Geolocation in the Browser: From Google Gears to Geolocation Sensors. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316538

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316538

1 HISTORY OF BROWSER GEOLOCATION

Geolocation has been available to developers implicitly through the mapping of Internet Protocol (IP) addresses or address blocks to known locations. There are numerous paid subscription and free geolocation databases with varying claims of accuracy that range from country level to state or city level, sometimes including ZIP/post code level. A popular free way to obtain IP-based location data was through Google's AJAX API Loader library [2], which provided an approximate, region-level estimate of a user's location through its `google.loader.ClientLocation` property. This API did not require users to install any client-side software. Implicit geolocation access is *not* subject of the paper, however, we list it for the sake of completeness and because it still plays an important role today, for example, for IP-based regional content blocking.

1.1 Google Gears Geolocation API

More accurate location data is available through Wi-Fi access point or cell tower data, as well as through Global Positioning Service (GPS) data. This data is accessible to native applications, however, initially, not via Web APIs. The Google Gears browser plugin [9] for the first time exposed exact device-based geolocation data to the browser. Its Geolocation API had two JavaScript methods: `getCurrentPosition()` made a single, one-off attempt to get a position fix, while `watchPosition()` watched the user's position over time, and provided an update whenever the position changed. Both methods allowed the developer to configure which sources of location information would be used. As a simple way to get an approximate position fix with low cost in terms of both network and battery resources, Gears also kept track of the best position fix obtained from these calls and made it available as the `lastPosition` property. To a contemporary JavaScript developer who may have never used the plugin, the code in Listing 1 still looks very familiar, even more than ten years later.

```
var geo = google.gears.factory
    .create('beta.geolocation');

function updatePosition(position) {
  alert('Current lat/lon is: ' + position.latitude + ','
      + position.longitude);
}

function handleError(positionError) {
  alert('Attempt to get location failed: ' +
      positionError.message);
}

geo.getCurrentPosition(updatePosition, handleError);
```

Listing 1: Google Gears API (2008)

From the start, user privacy was a major concern, and the plugin authors wrote in the Wiki [10]:

> "It must be clear to users when an application is using the Geolocation API. We could implement one or both of the following UI elements:
> (1) A separate dialog from the Gears security dialog to enable the Geolocation API. If the general-purpose dialog gave access to position data, it would be easy for users to forget they allowed access to Gears, or to fail to realize enabling Gears also exposes their position data.
> (2) Some persistent UI that indicates the Geolocation API is being used. For example, there could be a bar across the bottom of the browser with an icon of a globe or map. Perhaps this UI should be 'active' somehow, indicating that something is happening, so that the user cannot forget it is being used."

In the end, they decided on approach (1), which is still how location access is granted today, despite different underlying mechanics.

The similarity of the code in Listing 1 with code one would write today is no surprise. Gears implemented the at the time current editor's draft of the W3C Geolocation specification, which some of the authors have helped to define in collaboration with Microsoft, Mozilla, and others. The goal for Gears was to advance browser capabilities and eventually make itself a relict of time, which is what has happened.

1.2 Mozilla Geode

Mozilla Geode [1] was an experimental add-on that appeared slightly after Gears to explore geolocation in Firefox 3 ahead of the implementation of the final API in a future product release. Geode provided an early realization of the W3C Geolocation specification [14] so that developers could begin experimenting with enabling location-aware experiences. It included a single experimental proprietary geolocation service provider called Skyhook to map Wi-Fi signals to locations, so that any computer with Wi-Fi could get accurate positioning data. An interesting historical fact is that the ultimate plan for Firefox was that service providers and geolocation methods would be pluggable [1] to provide users with as many choices and privacy options as possible, which has not happened.

1.3 W3C Geolocation API

The W3C Geolocation API specification [14] defines a high-level interface to location information associated only with the device hosting the implementation, such as latitude and longitude. The API itself is agnostic of the underlying location information sources, and no guarantee is given that it returns the device's actual location. The design enables both "one-shot" position requests as well as repeated position updates, and also includes the ability to explicitly query the cached positions. As outlined above, the work is based on prior art in the industry in the form of Gears and Geode, including the work of Turner, Popescu, Sarver, Raskin, and others in the context of locationaware.org and geolocation in the Firefox browser [15].

Similar to its predecessors Google Gears and Mozilla Geode, the specification defines three methods: the `getCurrentPosition()` method asynchronously attempts to obtain the current location of the device. If the attempt is successful, the `successCallback` must be invoked with a new `Position` object, reflecting the current location of the device. If the attempt fails, the `errorCallback` must be invoked with a new `PositionError` object, reflecting the reason for the failure. The `watchPosition()` method returns an identifier that uniquely identifies a watch operation and then asynchronously starts that watch operation. This operation must first attempt to obtain the current location of the device. Similar to the "one-shot" request, if the attempt is successful, the `successCallback` must be invoked with a new `Position` object, reflecting the current location of the device, else, it must fail as described before. The watch operation then must continue to monitor the position of the device and invoke the appropriate callback every time this position changes. The watch operation must continue until the `clearWatch()` method is called with the corresponding identifier, which immediately stops the watch operation identified by the identifier. Listing 2 shows all methods in operation (see Listing 1 for comparison). This API is almost universally supported in browsers, Listing 1 shows the support situation at the time of writing.

```
function showMap(pos) {
  // Show a map centered at
  // (pos.coords.latitude, pos.coords.longitude)
}

function scrollMap(pos) {
  // Scrolls the map so that it is centered at
  // (pos.coords.latitude, pos.coords.longitude)
}

function handleError(error) {
  // Handles the error gracefully.
}

// One-shot position request
navigator.geolocation.getCurrentPosition(showMap,
     handleError);

// Request repeated updates
var watchId = navigator.geolocation.watchPosition(
    scrollMap, handleError);

function clearPositionWatch() {
  // Cancel the updates when the user clicks a button
  navigator.geolocation.clearWatch(watchId);
}
```

Listing 2: Geolocation API

1.4 W3C Geofencing API

The W3C Geofencing API [12] specification defined an API that should let Web applications set up geographic boundaries around specific locations and then receive notifications when the hosting device entered or left those areas, also referred to as geofences. The Geolocation Working Group was chartered to define a secure and privacy-sensitive interface for using client-side location information in location-aware Web applications and published an update to the Geolocation API [14] specification as a Recommendation excluding the geofencing feature, which was being worked on in the context of the separate Geofencing API specification, however, the work on it did not complete. While it would be possible to implement something similar to geofencing using the Geolocation API

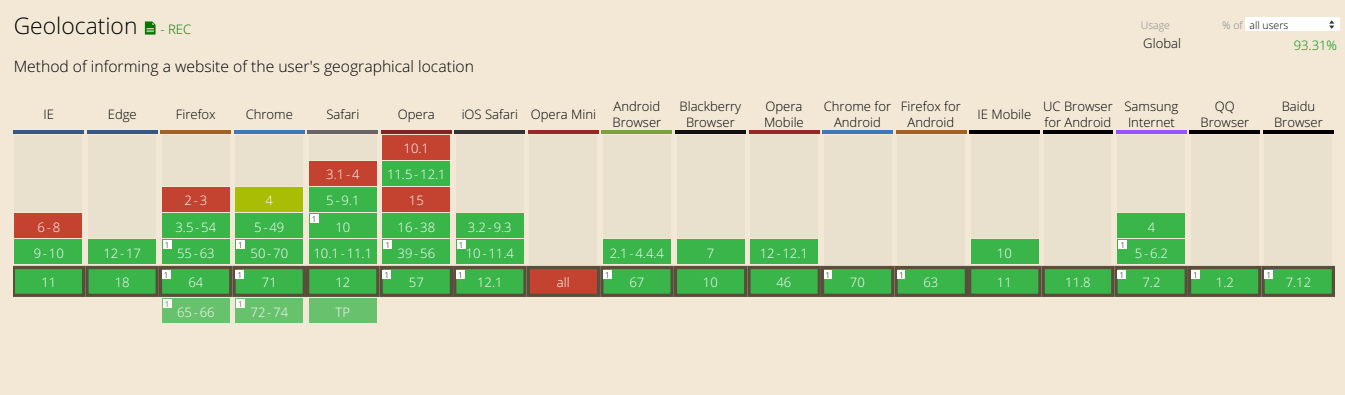


Figure 1: Browser support data for the Geolocation API (green means supported, red means not supported), data via https://caniuse.com/#feat=geolocation

(as long as the application is in the foreground), there were a few differences that made the proposed API look like a better choice, quoting directly from the specification [12]:

> "(1) Because of the limited API surface of the Geofencing API, user agents can implement the API in a more (power) efficient way than could be achieved by regularly checking the current geographic position with the Geolocation API.
> (2) The Geofencing API is built around Service Workers. This makes it possible for a webapp to receive notifications from the Geofencing API even after the user has closed the webapp."

It was not generally agreed on that Service Workers were the right technology choice,[1] as they may be killed by the user agent at nearly any time, which means one cannot expect `watchPosition()` to keep watching for extended periods, and likewise the method `getCurrentPosition()` could only work if the position was already available, but if it has to wait for a GPS update, the Service Worker is likely to be gone by the time an answer comes back.

2 CURRENT GEOLOCATION EFFORTS

After having covered the significant milestones in the history of geolocation in the browser, the second part of the paper introduces some of the current efforts, challenges, and use cases in the field.

2.1 Background Geolocation with Wake Locks

When a device that has either of the two Geolocation API methods `watchPosition()` or `getCurrentPosition()` running goes into stand-by mode (colloquially "goes to sleep"), or when the browser window is backgrounded, location reporting stops soon thereafter. This renders certain use cases like fitness run trackers or maps navigation that both require background geolocation entirely impossible. Wake Locks, initially introduced with the since discontinued Firefox OS [8], provided a way[2] to prevent the system from sleeping and to keep certain services running when the developer wanted to keep an invisible application continuing to use GPS. This could be done by requesting a `MozWakeLock` of the type `"gps"` and

[1] https://www.w3.org/2014/10/27-28-geolocation-minutes.html#day21

[2] https://developer.mozilla.org/en-US/docs/Archive/B2G_OS/API/Wake_Lock_API/Keeping_the_geolocation_on_when_the_application_is_invisible

```
// On main page
navigator.serviceWorker
  .register('serviceworker.js')
  .then((swRegistration) => {
    let region = new CircularGeofenceRegion({
      name: 'myfence',
      latitude: 37.421999,
      longitude: -122.084015,
      radius: 1000
    });
    let options = {
      includePosition: true
    };
    swRegistration.geofencing.add(region, options)
      .then(
        // If more than just a name needs to be stored
        // with a geofence, now would be the time to
        // store this in some storage
        (geofence) => console.log(geofence.id),
        (error) => console.log(error)
      );
});

// In serviceworker.js
self.ongeofenceenter = (event) => {
  console.log(event.geofence.id);
  console.log(event.geofence.region.name);

  // If this is not a geofence of interest anymore,
  // remove it now
  if (event.geofence.region.name !== "myfence") {
    event.waitUntil(event.geofence.remove());
  }
};
```

Listing 3: Geofencing API (conceived example)

```
try {
  wakeLockObj = await navigator.getWakeLock('system');
  wakeLockObj.addEventListener('activechange', () => {
    wakelock.textContent =
        `The ${wakeLockObj.type} wake lock is ${
        wakeLockObj.active ? 'active' : 'not active'}.`;
  });
} catch (err) {
  console.error('Could not obtain wake lock', err);
}

const toggleWakeLock = () => {
  if (wakeLockRequest) {
    wakeLockRequest.cancel();
    wakeLockRequest = null;
    return;
  }
  wakeLockRequest = wakeLockObj.createRequest();
};

const startTracking = () => {
  id = navigator.geolocation.watchPosition(success,
      error, opt);
};

trackButton.addEventListener('click', () => {
  toggleWakeLock();
  if (id) stopTracking();
  else startTracking();
});
```

Listing 4: Using the Wake Lock API for background geolocation tracking

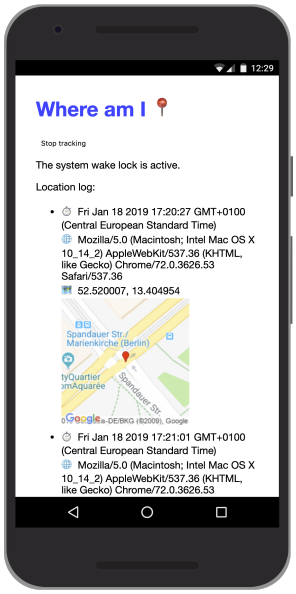


Figure 2: Web app "Where Am I" with Wake Lock for background location tracking (https://whereami.glitch.me/)

using it together with `watchPosition()`. App developers needed to be responsible and think carefully about whether they needed to keep the geolocation service on or not. The risk of claiming the lock was (and still is) that users may forget to close the app when they were done using it, which inevitably results in significantly increased battery use, apart from the privacy implications of potentially inadvertently continuing to share one's location.

Work on the concept of Wake Locks has been resumed in the form of the w3c Wake Lock specification [3]. With the at the time of writing implemented version of the specification, Wake Locks of the type `"system"` (currently there is no type `"gps"`) can be used together with `watchPosition()`. Listing 4 shows the relevant code excerpts of a Web application called "Where Am I", depicted in Figure 2, that implements this, simulating a simple run tracker that works with the device screen off while the user runs. The accompanying app "There Am I"[3] can be used to visualize users.

2.2 Geolocation Sensor

The w3c Geolocation Sensor specification [11] extends the `Sensor` interface defined in the w3c Generic Sensor API [20] to in turn define the new `GeolocationSensor` interface for obtaining the geolocation of the hosting device. The Generic Sensor API is a set of interfaces that expose sensor devices to the Web platform. It consists of the base `Sensor` interface and a set of concrete sensor classes (for example `GeolocationSensor`, `AmbientLightSensor`, `Accelerometer`, *etc.*) built on top. Having a base interface simplifies the implementation and specification process for the concrete sensor classes. The core functionality is specified by the base interface, and `GeolocationSensor` merely extends it with a tiny set of additional attributes and methods. The feature set of `GeolocationSensor` is similar to that of the Geolocation API [15], but it is surfaced through a modern API that is consistent across contemporary Sensor APIs, improves security and privacy, and is extensible. The API aims to be polyfillable[4] on top of the existing Geolocation API. As all recent APIs, instead of callbacks, Geolocation Sensor uses promises for "one-shot" or "one-and-done" operations [6].

Unlike with the previous Geolocation API, the Geolocation Sensor API does not have dedicated callbacks for "one-shot" versus continuous location requests,[5] but instead fires an event according to a frequency (in Hertz) that is used to calculate the requested sampling frequency for the associated platform sensor and to define the upper bound of the reporting frequency for the `GeolocationSensor` object. Listing 5 shows the API in practice, however, running it requires a polyfill, as currently there is no native implementation.

3 GEOLOCATION PRIVACY

Security and privacy issues for location-based services and geolocation-capable applications often revolve around designing a user interface such that users are informed about what an application is doing and have the ability to accept or decline [7]. Use cases like background geolocation or geofencing demand for a thorough re-evaluation of privacy-related questions. Similarly, some of the requirements that today we take for granted just did not exist when the Geolocation API was designed, like the strict requirement to be on a secure connection for using modern APIs.[6]

[3] https://whereami.glitch.me/ and https://thereami.glitch.me/

[4] https://github.com/w3c/sensors/blob/master/polyfills/geolocation.js

[5] `GeolocationSensor` does have a dedicate static operation `read()` for "one-shot"

[6] Internet Explorer up to version 11 even warned its users "You are about to view pages over a secure connection" when they navigated to a site that used the HTTPS protocol.

```
// Get a new geolocation reading every second
const geo = new GeolocationSensor({frequency: 1});
geo.start();

geo.onreading = () => console.log(
    `lat: ${geo.latitude}, long: ${geo.longitude}`);

geo.onerror = event => console.error(
    event.error.name, event.error.message);

// Get a one-shot geolocation reading
GeolocationSensor.read()
  .then(geo => console.log(
      `lat: ${geo.latitude}, long: ${geo.longitude}`))
  .catch(error => console.error(error.name));
```

Listing 5: Geolocation Sensor API

3.1 Insecure Origin Usage of Geolocation

As the Web platform is extended to enable more useful and powerful applications, it becomes increasingly important to ensure that the features which enable those applications are enabled only in contexts which meet a minimum security level. The Secure Contexts specification [21] describes threat models for feature abuse on the Web as follows:

> "Granting permissions to unauthenticated origins is, in the presence of a network attacker, equivalent to granting the permissions to any origin. The state of the Internet is such that we must indeed assume that a network attacker is present. Generally, network attackers fall into two classes: passive and active.
> **1. Passive Network Attacker:** A 'Passive Network Attacker' is a party who is able to observe traffic flows but who lacks the ability or chooses not to modify traffic at the layers which this specification is concerned with. [...]
> **2. Active Network Attacker:** An 'Active Network Attacker' has all the capabilities of a 'Passive Network Attacker' and is additionally able to modify, block or replay any data transiting the network. These capabilities are available to potential adversaries at many levels of capability, from compromised devices offering or simply participating in public wireless networks, to Internet Service Providers indirectly introducing security and privacy vulnerabilities [...], to parties with direct intent to compromise security or privacy who are able to target individual users, organizations or even entire populations."

In a rare case of breaking backwards compatibility and after a careful risk analysis, the concepts in [21] were applied to features that had already shipped in browsers and which did not meet the—new, not present at the time—requirements. Specifically, geolocation support was removed on insecure origins, motivated by the large privacy risk for end users of even passive attackers sniffing geolocation obtained from this API.

3.2 Feature Policy Integration

Feature Policy [4] defines a mechanism that allows developers to selectively enable and disable use of various browser features and APIs.

The Feature Policy integration in the Generic Sensor API is used to control access to sensors data for a frame. By default the `Sensor` objects (and therefore the `GeolocationSensor`) can be created only within a main frame or same-origin subframes, thus preventing cross-origin iframes from unsanctioned reading of sensor data. This default behavior can be modified by explicitly enabling or disabling of the corresponding policy-controlled features. Listing 6 illustrates granting geolocation data access to a cross-origin iframe, meaning that now `GeolocationSensor` objects can be created there.

```
<iframe src="https://3rd-party.com" allow="geolocation">
```

Listing 6: Allowing an iframe to use `GeolocationSensor`

3.3 Focused Area and Visibility State

Sensor readings are only available for active documents whose origin is same origin-domain with the currently focused area document and whose visibility state is `"visible"`. This is done in order to mitigate the risk of a skimming attack against the browsing context containing an element which has gained focus, for example when the user carries out an in-game purchase using a third party payment service from within an iframe. Similar to Subsection 2.1, these to be re-evaluated constraints currently limit use cases like fitness run trackers or maps navigation that require sensor readings in potentially non-active, non-visible documents, for example while the user chooses a music track to accompany their run or drive.

3.4 Permissions and Privacy Policy

Access to sensor readings are controlled by the w3c Permissions API [13]. User agents use a number of criteria to grant access to the readings. Note that while access to some sensors can be granted without prompting the user, `GeolocationSensor` always requires a prompt. Listing 7 shows the permissions flow that is required before any readings. In contrast, with the legacy Geolocation API, the user agent prompts the user automatically upon the first time they try to use either of `watchPosition()` or `getCurrentPosition()`.

```
navigator.permissions.query({name: "geolocation"})
  .then(({state}) => {
    switch (state) {
      case "granted":
        showLocalNewsWithGeolocation();
        break;
      case "prompt":
        showButtonToEnableLocalNews();
        break;
      default:
        break; // Don't do anything if permission denied
    }
  });
```

Listing 7: Asking for permission to use `GeolocationSensor`

The current permission prompt has the options to allow or to block the location access request (or, for what it is worth, to ignore it by closing the dialog). Interestingly, Raskin envisioned [15] a far more detailed permissions dialog, depicted in Figure 3, which had granular levels of permitted access, for example, to limit the

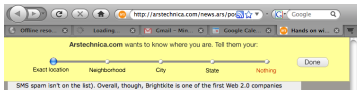


Figure 3: Aza Raskin's security prompt mock-up [15]

granularity to city or state level. However, this was never implemented, as despite being fully investigated, it was found to be nearly impossible to implement "fuzzing" of location data [19].

With use cases like run trackers, it would additionally be of interest to limit access to background geolocation data to a given time period, for example, the duration of a one hour run. This is currently not possible, and the Permissions specification [13] is intentionally vague regarding the storage of permissions, leading to some user agents (*e.g.*, Chrome) to persist decisions, others (*e.g.*, mobile Safari) to ask every time, and yet others (*e.g.*, desktop Safari or Firefox) to optionally remember them (time-limited for Safari).

A notable privacy discussion was happening between the IETF GEOPRIV and the W3C Geolocation Working Groups around the question of whether or not a privacy policy should be included in the Geolocation API itself, or rather be addressed as a set of recommendations and requirements in the specification. The Geolocation Working Group concluded[7] that privacy protection was better handled as part of a more generic privacy and security framework for device access. Privacy and security threats now take a significant amount of space in the Generic Sensor specification [20].

4 FUTURE WORK AND CONCLUSION

Future work will happen in several areas and touch upon both technological as well as user privacy related aspects. In the following paragraphs, we present some of them, but note that especially the permission aspects reach a lot further than geolocation.

We will continue our experiments with Wake Locks [18] and how they can interact with Geolocation Sensor, with a special focus on sensor readings in non-focused and non-visible states (Subsection 3.3). Further, after having explored use cases around background geolocation and geofencing [11] in more detail, we will resume our work on Geofencing [12] and investigate integration of this feature with Geolocation Sensor [11], possibly via the extensibility mechanism defined in Service Workers [17] that makes them extensible from other specifications through a functional event by extending the `ExtendableEvent` interface.

User privacy is central to all sensor readings, and particularly to geolocation. As outlined in Subsection 1.1, from the start it was a major concern and especially with new background geolocation and geofencing use cases, it remains an issue of high importance. In [16], Russell and Nattestad have explored permission dialog abuse, annoyance, and fatigue of users; and according to their opinion (which we support), "today's permissions model and API surface area should be heavily revised [...] to enable more flexible permission policies" in a more uniform Permissions API [13]. Some ideas include supporting both time-limited as well as one-off requests and permission grants, but also extending Web Application Manifest [5] to allow sites to identify to the runtime a maximum set of permissions, and turn down permission requests not explicitly included in the Web Application Manifest, with installability [5] of Web apps potentially considered an additional trust factor.

Concluding, in this paper we have first provided an overview of the historical development of geolocation in the browser, and given an outlook on current and future efforts, challenges, and use cases in the field. The importance of user privacy is ever-increasing, and in our work on geolocation, background geolocation, and geofencing, we keep privacy as a top priority. While many privacy concerns such as unsanctioned third-party access and fingerprinting can now be mitigated by technical measures, we continue to work with the privacy research community to ensure the Web remains a both capable and safe platform today and in the future by carefully assessing every new proposed feature for possible privacy threats.

[7] https://www.ietf.org/lib/dt/documents/LIAISON/file727.txt

ACKNOWLEDGEMENTS

We are thankful to *Doug Turner* for his invaluable help with getting the historical facts right and to *Alex Russell* for reviewing the paper.

REFERENCES

[1] Chris Beard. 2018. *Introducing Geode.* Blog Post. Mozilla Labs, https://blog.mozilla.org/labs/2008/10/introducing-geode/.
[2] Steve Block. 2008. *Two new ways to location-enable your web apps.* Blog Post. Google Developers, https://developers.googleblog.com/2008/08/two-new-ways-to-location-enable-your.html.
[3] Ilya Bogdanovich, Andrey Logvinov, and Marcos Cáceres. 2017. *Wake Lock API.* Candidate Recommendation. W3C, https://www.w3.org/TR/wake-lock/.
[4] Ian Clelland. 2019. *Feature Policy.* Editor's Draft. W3C, https://w3c.github.io/webappsec-feature-policy/.
[5] Marcos Cáceres, Kenneth Rohde Christiansen, Mounir Lamouri, Anssi Kostiainen, Rob Dolin, and Matt Giuca. 2018. *Web App Manifest, Living Document.* Working Draft. W3C, https://www.w3.org/TR/appmanifest/.
[6] Domenic Denicola. 2018. *Writing Promise-Using Specifications, Finding of the W3C TAG, 9 November 2018.* Technical Report. W3C, https://www.w3.org/2001/tag/doc/promises-guide/.
[7] Nick Doty and Erik Wilde. 2010. Geolocation Privacy and Application Platforms. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS (SPRINGL '10).* ACM, New York, NY, USA, 65–69. https://doi.org/10.1145/1868470.1868485
[8] Ben Francis. 2017. *The Story of Firefox OS.* Blog Post. Medium.com, https://medium.com/@bfrancis/the-story-of-firefox-os-cb5bf796e8fb.
[9] Google Gears. 2008. *Geolocation API.* Documentation. Google, https://web.archive.org/web/20080902073745/http://code.google.com:80/apis/gears/api_geolocation.html.
[10] Google Gears. 2008. *GeolocationAPI.* Wiki. Google, https://web.archive.org/web/20080901235406/http://code.google.com:80/p/gears/wiki/GeolocationAPI.
[11] Anssi Kostiainen, Thomas Steiner, and Marijn Kruisselbrink. 2018. *Geolocation Sensor.* Working Draft. W3C, https://www.w3.org/TR/geolocation-sensor/.
[12] Marijn Kruisselbrink. 2017. *Geofencing API.* Working Group Note. W3C, https://www.w3.org/TR/geofencing/.
[13] Mounir Lamouri, Marcos Cáceres, and Jeffrey Yasskin. 2017. *Permissions.* Working Draft. W3C, https://www.w3.org/TR/permissions/.
[14] Andrei Popescu. 2016. *Geolocation API Specification 2nd Edition.* Recommendation. W3C, https://www.w3.org/TR/geolocation-API/.
[15] Aza Raskin. 2010. *Geolocation in Firefox and Beyond.* Blog Post. Aza on Design, http://www.azarask.in/blog/post/geolocation-in-firefox-and-beyond/.
[16] Alex Russell and Thomas Nattestad. 2018. *Permissions Workshop.* Position Paper. W3C, https://www.w3.org/Privacy/permissions-ws-2018/papers/thomas-nattestad.pdf.
[17] Alex Russell, Jungkee Song, Jake Archibald, and Marijn Kruisselbrink. 2017. *Service Workers 1.* Working Draft. W3C, https://www.w3.org/TR/service-workers/.
[18] Thomas Steiner. 2018. *Experimenting with the Wake Lock API.* Blog Post. Medium.com, https://medium.com/dev-channel/experimenting-with-the-wake-lock-api-b6f42e0a089f.
[19] Martin Thomson. 2011. *Obscuring Location.* Internet Draft, GEOPRIV. IETF, https://tools.ietf.org/id/draft-thomson-geopriv-location-obscuring-03.html.
[20] Rick Waldron, Mikhail Pozdnyakov, and Alexander Shalamov. 2018. *Generic Sensor API.* Candidate Recommendation. W3C, https://www.w3.org/TR/generic-sensor/.
[21] Mike West. 2016. *Secure Contexts.* Candidate Recommendation. W3C, https://www.w3.org/TR/secure-contexts/.

Hashtag Usage in a Geographically-Local Microblogging App

Helge Reelfs, Timon Mohaupt, Oliver Hohlfeld
RWTH Aachen University

Niklas Henckell
The Jodel Venture GmbH

ABSTRACT

This paper studies for the first time the usage and propagation of hashtags in a new and fundamentally different type of social media that is *i)* without profiles and *ii)* location-based to only show nearby posted content. Our study is based on analyzing the mobile-only Jodel microblogging app, which has an established user base in several European countries and Saudi Arabia. All posts are user to user anonymous (i.e., no displayed user handles) and are only displayed in the proximity of the user's location (up to 20 km). It thereby forms local communities and opens the question of how information propagates within and between these communities. We tackle this question by applying established metrics for Twitter hashtags to a ground-truth data set of Jodel posts within Germany that spans three years. We find the usage of hashtags in Jodel to differ from Twitter; despite embracing local communication in its design, Jodel hashtags are mostly used country-wide.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Networks** → *Online social networks.*

KEYWORDS

Anonymous Messaging; Location Based Messaging; User Behavior and Engagement; Information Diffusion; Hashtag

ACM Reference Format:
Helge Reelfs, Timon Mohaupt, Oliver Hohlfeld and Niklas Henckell. 2019. Hashtag Usage in a Geographically-Local Microblogging App. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316537

1 INTRODUCTION

Social media has become a popular and ubiquitous tool for consuming and sharing digital content (e.g., textual or multimedia). This sharing leads to information propagation and spreading across users and even across different networks [20]. Understanding this propagation has thus motivated research studies to investigate the dynamics of information adoption, spreading, and (complex) contagion of information [2, 4–6, 12, 14, 17, 19], e.g., in the form of memes. A widely studied platform in this regard is the microblogging service Twitter that enables users to reach a global audience and for which sampled post data is available via APIs. Analyzing the post contents' (e.g., included memes) is, however, a very challenging application of natural language processing. Since users often self-classify their posts by adding hashtags to ease retrieval, analyzing hashtags is a promising proxy measure for analyzing memes or post contents. This has resulted in metrics to analyze hashtags and thereby valuable insights into their spreading behavior [6].

New *location-based* and user to user *anonymous* microblogging services complement classical social media platforms and their design differences open the question if classical observations on information spreading are still applicable. One emerging platform in this regard is the *Jodel* mobile-only microblogging app. Launched in 2014, it has been widely adopted in several European countries and Saudi Arabia. Like Twitter, Jodel enables users to share short posts of up to 250 characters long and images, i.e., microblogging. Unlike Twitter and other classical social media platforms, Jodel *i)* does not have user profiles rendering user to user communication anonymous, and *ii)* displays content only in the proximity of the user's location, thereby forming local communities. Despite the emerging use of such platforms, little is known on how their key design differences impact information propagation.

In this paper, we present the first study on information spreading in such an emerging platform by investigating the hashtag propagation in Jodel as a prominent application in this space. We take a detailed look on hashtag propagation through the lens of a platform operator by having the unique opportunity to analyze data provided by Jodel for messages posted in Germany from September 2014 to August 2017. This longitudinal data set enables us to study how this key design pattern of forming local communities by only displaying content to nearby users influences the hashtag usage and compares to the global counterpart Twitter. Our study is based on using established metrics designed to capture the spatial focus and spread of Twitter hashtags [6] to Jodel. We show that these metrics can be applied to the temporal dimension to cover the spread of hashtags in time, enabled by our longitudinal observation period. We further study similarities in hashtag usage between cities and their spacial impact—finding that larger cities/communities influence the smaller ones. The correlation of spatial and temporal metrics reveal that hashtags can be grouped into four different hashtag classes distinguished by their spatial and temporal extent. In the last step we show that these groups are distinguishable by machine learning models, informed by manual labeling of 450 most frequently used hashtags. Our main contributions are as follows:

- We provide the first comprehensive study of hashtag usage in a local user to user anonymous messaging app. We find that Jodel's popular hashtags are used country-wide, whereas less popular hashtags tend to be more local.
- We show that classical metrics capturing the spatial propagation can be applied to the temporal domain. By applying these metrics, we see that popular hashtags are used over the long-run, while less popular hashtags tend to be more short-lived.
- We show that the used hashtags can be grouped into four classes by their spatial and temporal extent. We further show that these

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316537

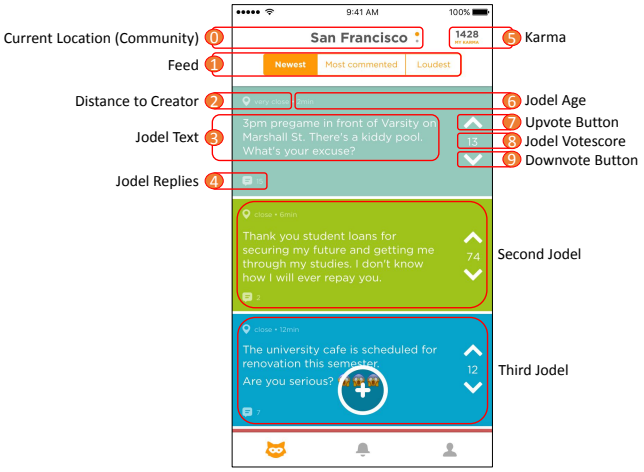


Figure 1: Jodel iOS mobile application.

four groups can be learned by statistical models with high accuracy, based on comparing five different classifiers (k-nearest neighbour, regression trees, naive bayes, LDA, ZeroR). Thus, statistical methods can distinguish between different meme types found in Jodel.

Paper structure. We introduce Jodel in Section 2 and discuss related work in Section 3. Section 4 introduces our Jodel dataset to which we apply established hashtag propagation metrics in Section 5. In Section 6, we show that our findings can be leveraged to classify hashtags automatically. We conclude the paper in Section 7.

2 JODEL - LOCAL MESSAGING APP

Jodel[1] is a mobile-only messaging application (main-screen shown in Figure 1). Unlike classical social media apps, it is location-based and establishes local communities to the users' location ⓪. Within these communities, users can post both images and textual content of up to 250 characters length ③ (i.e., microblogging) *anonymous* to other users—and reply to posts forming discussion threads ④. Posted content is referred to as "Jodels" and are colored randomly ③. These posts are only displayed to other users within close (up to 20 km) geographic proximity ②. This ability to only consume local content is absent in typical social networks (e.g., Twitter) that enable global communication and thus makes the study of information spread interesting.

All communication is *anonymous to other users* since no user handles or other user-related information are displayed. Only *within* a single discussion thread, users are enumerated and represented by an ascending number in their post order. There are three different content feeds ①: *i) newest* showing the most recent threads, *ii) most commented* showing the most discussed threads and *iii) loudest* showing threads with the highest voting score (cf. later). Additionally, users can subscribe to thematic channels. Each post can contain hashtags and the app enables to display further local posts with the same hashtag by clicking on a hashtag in a post.

[1]Jodel, German for yodeling, a form of singing or calling.

Jodel employs a community-driven filtering and moderation scheme to avoid adverse content. For any social network or messaging app, community moderation is a key success parameter to prevent harmful or abusive content. The downfall of the Jodel-alike YikYak anonymous messaging application highlighted that unsuccessfully preventing adverse content can seriously harm it [10]. In Jodel, content filtering relies on a distributed voting scheme in which every user can increase or decrease a post's vote score by up- (+1) ⑦ or downvoting (-1) ⑨, i.e., similar to StackOverflow. Posts reaching a cumulative vote score ⑧ below a negative threshold (e.g., -5) are not displayed anymore. Depending on the number of vote-contributions, this scheme filters out bad content while also potentially preferring mainstream content. As a second line of defense, Jodel employs community moderators who decide on removing reported posts.

3 RELATED WORK

Our paper relates to three main areas within research: *i)* general meme spread modelling, *ii)* the use case of microblogging, e.g., Twitter, and *ii)* others; which we will discuss next.

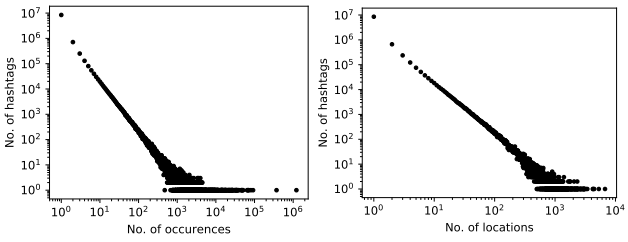
Spreading & contagion models. A classical approach to study information diffusion is applying spreading models. Epidemic models have been applied to memes, where a meme can *infect* people by coming in contact with it (SIR models)—possibly extended with mechanics for *recovery* (SIRS models), e.g., in [17, 19]. Although these approaches model the growth of hashtag popularity well, most fail to map the typical power-law decay [11]. Their application to hashtags is further limited by requiring an infection time, i.e., when a user learns about a hashtag. Passive information consumption such as reading is typically not included in most social network data.

Twitter. The study of hashtag usage and diffusion mostly targets Twitter given its popular use of hashtags and ability to geotag posts. Although Twitter has no boundaries regarding distance (i.e., unlike Jodel), cities closer to each other share more hashtags, supported by an analysis of the Twitter trending topics in [5]. The authors find three clusters of hashtag similarity across the biggest cities in the US and speculate that the spread is related to airports. To study non-stationary time series of hashtag popularity, [14] applies a statistical measure originally used for neuron spike trains to hashtags. It is capable of giving information on how regularly hashtags are used. They find that low to mediocre popular Twitter hashtags are on average rather bursty, while extremely popular ones are posted more regularly. The influence of content (e.g., politics, music, or sports) on the hashtag adoption is studied in [12]. The authors find that especially political hashtags are more likely to be adopted by a user after repeated exposure to it than hashtags of other topics.

To capture the spatio-temporal dynamics of Twitter hashtags, *focus*, *entropy*, and *spread* were proposed as metrics [6]. By applying these metrics to Twitter, the authors find hashtags to be a global phenomenon but the distance between locations to constraint their adoption. We will use these metrics to study Jodel and we extend them with a temporal dimension within our analysis. To study the how cities impact each other regarding hashtag adoption, [6] also proposed a spatial impact metric to capture the similarity of hashtag uses in two cities—a metric that we will adopt likewise. They show that the biggest influencers were big cities with large user bases.

Table 1: Hashtag dataset statistics. The data ranges from the application start in late 09/2014 up to beginning of 08/2017.

Metric	#Entries	Description
Hashtag Uses	41, 038, 733	# of hashtags occurrences
Hashtags	13, 110, 573	# of different hashtags
	11, 092, 360	# of hashtags used only once
Messages	26, 955, 008	# of messages that contained hashtags
Users	1, 240, 404	# of users posting contents with hashtags
Locations	6, 830	# of different posting locations/cities



(a) Hashtag distribution **(b) Location distribution**

Figure 2: These figures show *a)* the hashtag distribution w.r.t occurrences and the corresponding amount, *b)* the location distribution w.r.t occurrences for a hashtag and the corresponding amount—both distributions are heavy-tailed.

Other platforms. Besides Twitter, few studies consider other platforms. The sharing cascades in Facebook are studied in [4]. Similar cascades are found by studying how the blogosphere and the news media influence each other [8]. Memes do not have to be in the form of images or text, but can also be videos–as such, e.g., [18] studies the diffusion of memes on Youtube.

Other works focused on the influence of events in terms of the spreading behavior. E.g., [1, 7] used statistical classifiers on contextual features to distinguish between memes and events. Researchers have also tried to detect events, e.g., by analyzing the Twitter stream [9, 16] and inferring where an event happens [15]. There were also efforts to detect earthquakes and estimating the epicenter in real-time [13]. Also, user positions can be at least vaguely estimated as shown in [3].

We complement these works by studying the hashtag usage and diffusion on Jodel. Its property to only display posted content to nearby users differentiates Jodel from other studied social networks that disseminate content globally (e.g., Twitter or Facebook). It thus *might*—and as we will see: *will*—feature a fundamentally different spreading behavior.

4 DATASET DESCRIPTION AND STATISTICS

The Jodel network operator provided us with *anonymized* data of their network. This obtained data contains post, user and interaction metadata and message contents created within Germany only. It spans multiple years from the beginning in September 2014 of the network up to August 2017. The dataset only includes infromation users have publicly posted and thus visible to all other Jodel users. Structurally, our available dataset is built up from three object categories: interactions (about 400 M records), content (about 285 M records), and users (about 900 k records). The location of each post (and thus each hashtag) is available on a *city-level granularity*.

Hashtags. We have extracted hashtags from the message contents by applying a regular expression matching a '#' followed by any amount of alphanumeric characters (including German umlauts and Eszett), dots, dashes or underscores. This resulted in a total amount of about 41 M hashtag uses within 26 M different messages and 13 M different hashtags. These messages where created by 1.2 M users having posted in about 7 k different locations.

Within the set of hashtags, we observe that 11.1 M are only used once. This leaves about 2 M hashtags that have been used multiple times, i.e., ≥ 2, and therefore are suited for our hashtag propagation analysis at all. After manual sample screening, the predominant reason for this huge amount of hashtags occurring only once is that on Jodel, they are often used as a unique stylistic feature, support content, or are misspelled reuses—in contrast to a self-categorization that might be expected.

5 JODEL HASHTAG USAGE AND SPREAD

In this section, we analyze the spread and propagation of content in Jodel by using hashtags as a proxy measure. That is, we leverage the user's ability to tag posts with hashtags to relate to topics, add categories or metadata to posts. Although hashtags are sometimes used as a rather stylistic feature (e.g., by using numbers as hashtags to link multiple character limited posts together), more popular ones overall reasonably capture topics and memes in the posts.

We will see that some hashtags are specific to the Jodel platform and very local possibly due to its location-based design. Beginning our analysis in this Section with a study of hashtag popularity, we follow this up with their spatial and temporal spreading extent. We lastly study the hashtag usage in different cities and how they influence the hashtag adoption.

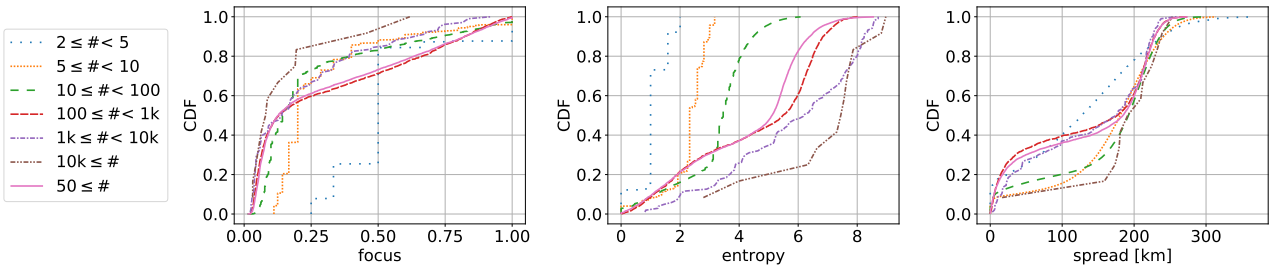
5.1 Overall Hashtag Use

Our data set consists of 27 M posts with hashtags. We overall find 41 M occurrences of 13 M unique hashtags of which only 2 M are used multiple times (cf. Table 1).

Popularity. We begin by studying the hashtag popularity. Figure 2a shows the distribution of a hashtag's occurrence (x-axis) vs. the corresponding amount of unique hashtags in our dataset (y-axis) on a log-log scale. We observe that the vast majority of hashtags are only used few times. The distribution is heavy-tailed and of similar shape, as observed in Twitter [6].

Location distribution. We next study how many hashtags (y-axis) are used in how many locations (x-axis) in Figure 2b. We see that not only the occurrences per hashtags is heavy-tailed, but also their geographic spread. These results are also very similar to Twitter [6].

Findings. We find most hashtags are being used only very few times. The hashtag usage follows a heavy-tailed distribution, which also holds true for the number of different locations in which they occur. That is, only a few hashtags are heavily popular and used in many locations—others to a lesser extent, or not.



(a) Spatial focus CDF **(b) Entropy CDF** **(c) Spread CDF**

Figure 3: Spatial hashtag metrics: focus, entropy and spread (left to right). All figures follow the partitioning by hashtag occurrences shown left. *a)* The more popular a hashtag is, the more unfocused it gets. Most hashtags show a very low focus, while very unpopular hashtags naturally tend to be more focused. *b)* Likewise, less used hashtags naturally can only be used in few locations, while more popular hashtags are used in many locations. *c)* Some hashtags on average span geographically only up to ±50 km, whereas most are used all over the country.

5.2 Spatial Properties of Jodel Hashtags

We next study spatial properties of Jodel hashtags, e.g., if a certain hashtag only occurs in a local community or over which geographic distance the usage of a countrywide hashtag is spread. To capture these spatial properties, we use three hashtag metrics originally proposed for Twitter: *focus*, *entropy*, and *spread* [6]. These metrics enable us to judge if content diffusion in Jodel actually is—due to its design—indeed *more local* than a comparable microblogging platform without geographical communities, like, e.g., Twitter.

Data filtering. We restrict our set of hashtags by only considering hashtags that occurred first in 2016 or later. This way, we focus on a time in which the app has an established user base in Germany.

Focus. The focus metric captures how locally or globally (i.e., in our case countrywide) focused the use of a hashtag is [6]. To achieve this, the set of hashtags and the set of locations are defined as H and L, respectively, of which for a given hashtag $h \in H$ and location $l \in L$, O_l^h is the set of occurrences of h in l. Then, the probability of observing a hashtag h in a location l is defined as:

$$P_l^h = \frac{|O_l^h|}{\sum_{m \in L} |O_m^h|}$$

The focus location of a hashtag is defined as the location with most occurrences of that hashtag and further provides a fraction of the occurrences in the focus location compared to the number of overall occurrences. It is defined as $F^h = \max_{l \in L} P_l^h$. Then, the focus for hashtag h is defined as a tuple of the focus location $l_f = F^h$ and its probability $P_{l_f}^h$. Hashtags only popular in a few cities will have a higher focus, whereas globally popular hashtags will have a lower one. A limitation of the focus metric is that it provides information only about one single location, but nothing about the distribution.

We show the focus distribution of hashtags in Figure 3a, where a series represents a CDF for a set of hashtags partitioned by their occurrence. As the hashtags are subject to popularity, i.e., usage frequency, these partitions define different log-based groups within out dataset (cf. Figure 2a). Our observation is that the focus distribution is skewed towards low focus values regardless of hashtag occurrences. That is, 60% of all hashtags that occur ≥ 5 times have a focus of ≤ 0.25. This means that from all occurrences of such a hashtag, only 25 % occur in its most popular city, whereas the remaining 75 % of the hashtag occurrences is in other cities. Therefore, the focus distribution indicates that the usage of most hashtags is not focused on a single city but is rather spread over multiple cites. Further, the observed skew within the distributions towards low focus values differs from hashtag usage observations in Twitter in which the hashtags' focus was uniformly distributed [6]. The prevalence of low focus values is unexpected and interesting; the design of the App to only display nearby posts could have caused a skew towards high focus values, in which the usage of most hashtags would be more concentrated. This, however, is not the case.

Entropy. The entropy metric captures in how many locations a hashtag is used [6]. For a hashtag h, it is defined as:

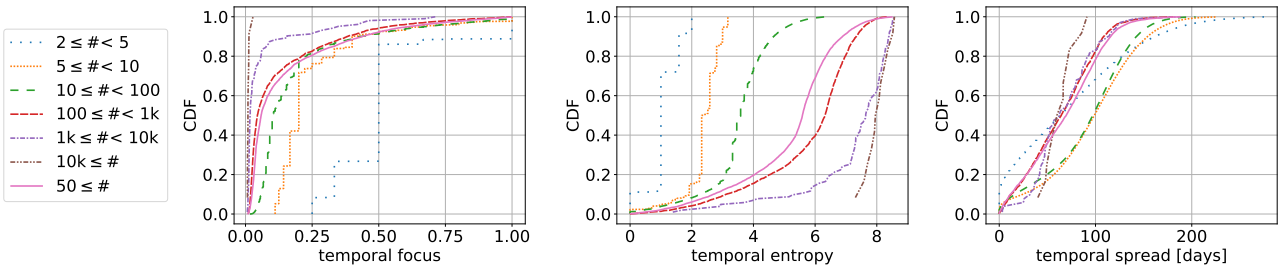
$$E^h = -\sum_{l \in L} P_l^h \log_2 P_l^h$$

This metric defines the minimum number of bits required to represent the amount of a hashtag's locations it has spread to. The higher the diffusion of a hashtag, the higher its entropy; i.e., the entropy defines the number of locations a hashtag occurred in by the power of 2. For more often used hashtags, both entropy and focus are resistant to small changes in the data (e.g., single occurrences in another ten locations).

Similar to the focus, we show the entropy distribution as CDFs for hashtags likewise partitioned by occurrences in Figure 3b. We observe that only a negligible number of hashtags is used in a single city (entropy 0). Looking into the different partitions, we identify that less popular hashtags clearly tend to a smaller entropy. However, for the more popular hashtags having at least 50 occurrences, more than 60 % of the hashtag occurrences are in ≥ 16 cities (entropy 4). As already indicated by the focus distribution, the usage of most hashtags is thus not concentrated to a single city only but spread over multiple cities. In summary, the hashtag usage shows a trend to higher entropy values with an increased number of occurrences; the more popular a hashtag is, there more it is spread across different cities, which supports our findings for the focus.

Spread. To obtain information about the geographical expansion, we can use the spread metric defined as the mean distance of the geographic midpoint of the set of hashtag occurrences [6]:

$$S^h = \frac{1}{|O^h|} \sum_{o \in O^h} D(o, G(O^h))$$



(a) Temporal Focus CDF **(b) Temporal Entropy CDF** **(c) Temporal Spread CDF**

Figure 4: Temporal hashtag metrics: temporal focus, temporal entropy, and temporal spread (left to right) are temporal adoption to the spatial counterparts (cf. Section 5.2). All figures follow the partitioning by hashtag occurrences shown left. *a)* The temporal focus decreases with hashtag popularity, i.e., they become used over longer time periods. *b)* This finding is supported by the temporal entropy showing that more popular hashtags are more widespread across multiple dates. *c)* The temporal spread indicates a possible distinction between a smaller set of short-lived hashtags and a large set of long-lived hashtags.

where D is the distance in kilometers and G is the weighted geographic midpoint. As on our scale (Germany), the spherical shape of the Earth is only of minor importance, we use the weighted average latitude and longitude as the midpoint. A spread of 50 km thus means that the average usage of a hashtag occurs within ±50 km.

We show the spread distribution again as CDFs of partitions by occurrences in Figure 3c. The distributions reveal that there are three groups of hashtags: *i)* Only rarely used hashtags (≤ 5 occurrences) show a rather linear spread, *ii)* More frequently used hashtags (5..100 occurrences) show a slight bimodal distribution as they either have a small spread up to 50 km, or most of them show a rather big spread > 150km. The same holds true for hashtags that are heavily used. *iii)* Hashtags that are used often, but do not belong to the heavy tail, strengthen the bimodal observation as about 40 % only have an up to 50 km, whereas most others are spread wider.

We note that higher spreads are likely the value a Germany-wide hashtag may achieve. While there is no (known) comparable analysis for Twitter or similar platforms, we conclude that the lower-spread hashtags are most probably an implication of Jodel's nature building location-based communities. I.e., there are hashtags that are used in a geographically restricted area at small distances.

Findings. We observe that most hashtags in Jodel are used rather countrywide, i.e., their usage does not concentrate on single cities and spreads over larger geographic distances. This is unexpected since the design of Jodel to form local geographic communities could also result in a more geographically focused usage of hashtags. However, while most hashtags are used rather globally, up to 40 % have a local spread of ±50 km and thus are a potential consequence of Jodels' design.

Twitter Comparison. A direct comparison to [6] can be made within our series of hashtags at least having 50 occurrences (pink solid lines). While the *focus* CDF for Twitter hashtags is rather linear with the exception of 20 % having focus 1, the focus on Jodel is distributed in an opposite fashion. That is, 60 % of Jodel hashtags (≥ 50 occurrences) tend to be non-focused below a value of 0.25, but are likewise equally distributed above—having almost no hashtags with focus 1. As for the *entropy*, most hashtags on Twitter are used very locally, which can only be observed for least popular hashtags on Jodel—many more popular hashtags are used across the country. Similarly, the *spread* on Twitter is either local for few hashtags, but then increases linearly, which is identical for the least and heavily popular hashtags on Jodel—others show a pronounced bimodal distribution between local and countrywide scope.

5.3 Temporal Properties of Jodel Hashtags

We are next interested in studying how hashtags develop over time (e.g., gain in popularity). This is possible given our longitudinal data set. Therefore, we adopted focus, entropy, and spread for our temporal analysis. Instead of locations as in our spatial analysis, we use the creation time of a hashtag's post (grouped to days for focus and entropy) for each hashtag occurrence. The grouping to days makes sense due to limited content presence within the usually highly dynamic Jodel feeds for larger communities.

Temporal Focus. We show the temporal focus distribution as CDFs partitioned by hashtag occurrences in Figure 4a. Recall that the temporal focus now defines the probability of a hashtag to be used on its most popular day, i.e., a temporal focus of 1 indicates that a hashtag is exclusively used on a single day whereas a focus of near 0 would suggest a spread over the entire observation period. We observe that about 80 % hashtags have a low temporal focus ≤ 0.25, suggesting that their lifetime is not focused on a single point in time. The more popular they become, the temporal focus decreases, i.e., they remain popular over time. However, least popular hashtags tend to a higher temporal focus in comparison. In summary, there are almost no hashtags focused to a single day. For those that are being used only a few times, this implicates random re-use that is probably not correlated, whereas popular hashtags are used throughout the observation period.

Temporal Entropy. The temporal entropy defines the number of days on which a hashtag is used. We show its distribution as CDFs partitioned by hashtags occurrences in Figure 4b. We observe that only a negligible amount of hashtags are used on exactly one day (entropy 0). Except for the only rarely used hashtags, more than 90 % occurrences have an entropy above 2, i.e., they were used on more than 4 (2^2) days. Further, the higher the occurrences (popularity) of a hashtag, the higher the entropy. This indicates that popular hashtags are used for longer time periods.

Temporal Spread. The temporal spread defines the average time period in days in which a hashtag is used. For example, a temporal spread of 50 days means that the average usage period of a

hashtag is ±50 days (past & future) from the temporal weighted midpoint. We show the distribution of the temporal spread as CDF again partitioned by hashtag's occurrences in Figure 4c. We observe that the temporal spread is distributed equal (linear CDF) across all partitions. However, the activity period is again influenced by the popularity of a hashtag; the more popular a hashtag is, the higher is the temporal spread. The presented series that only include hashtags with very few uses depict a large set of hashtags with a temporal spread of more than 100 days—the significant skew towards a larger spread strengthens our belief that such hashtags occur independently from each other (cf. temporal focus).

Findings. Popular hashtags in Jodel are seldomly a flash in the pan but are mostly used over extended time periods. In particular, the more popular a hashtag is, the longer and frequent its usage period becomes, whereas less popular ones rather occur independently from each other. This is interesting since the Jodel app provides—unlike Twitter—only limited functionality to search for hashtags as hashtags may only be clicked when seen in a post, i.e., for a purposeful re-use it must be known.

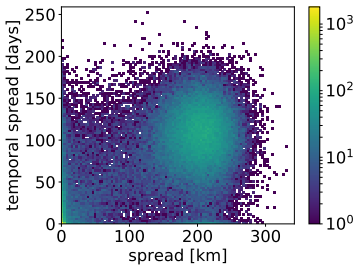
5.4 Spatial vs. temporal dimensions

Having analyzed the spatial and temporal dimensions in isolation, we are now interested in how they correlate. For example, hashtags that occur in one geographic area have a low spatial spread, but can be active over a short or longer timespan as indicated by the temporal spread. Therefore, we focus on correlating the spatial and temporal spread and omit other metrics since they provide a similar picture. Figure 5a shows the spatial spread on the x-axis and the temporal spread on the y-axis of all hashtags having at least 30 occurrences since 2016. The hashtags can roughly be clustered into four groups as shown in Figure 5b. *i)* A temporal spread of 100 days and a spatial spread of 250 km (long-lived and countrywide). We would expect countrywide hashtags that are statements and also memes in this group, as both kinds are often spread out on the landscape and rather long-lived. *ii)* Located around a spatial spread of 250 km, but the temporal spread is only a few days (short-lived and global). Hashtags in this group are, for example, about countrywide events. Also, some memes that are short-lived could be in that group. *iii)* Spread around 0 to 30 km and temporal spread of 0 to 70 days (long-lived and local). Here, we would expect hashtags about phenomena that are particularly local due to the community structure of Jodel. *iv)* Short-lived and local hashtags. This group can involve for example local events. We will base our content classification of hashtags in Section 6 on these identified groups.

Findings. The correlation of spatial and temporal spread clusters the hashtags into four groups, identified by long-lived vs. short-lived and countrywide/global vs. local spread. That is, there are some long-lived and short-lived countrywide hashtags, while we also identify long- and short-lived local hashtag occurrences.

5.5 Influence and Similarity of Cities

We have seen that some hashtags occur rather locally, which is an essential aspect of the Jodel application. We have also seen that many hashtags spread through many Jodel communities. Therefore, we next want to examine how much communities influence each other in the sense of causing other cities to adopt a hashtag. We are particularly interested in which cities source and popularize trends before others adopt them.



(a) The combined spatial (x-) and temporal (y-axis) spread shown as a heatmap of hashtags occurrences (z-axis). We observe clusters: *i)* countrywide long-lived hashtags, *ii)* countrywide short-lived hashtags, and *iii)* Both, local short- and long-lived hashtags.

temporal	local	global
long-lived	local phenomena	other memes
short-lived	local event	events

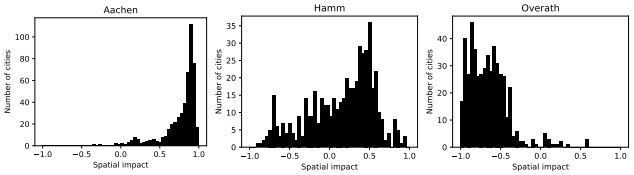
spatial

(b) Identified classes of hashtags within the Jodel platform according to the spatial and temporal spread metric.

Figure 5: The correlation between spatial and temporal spread. *a)* describes the number of hashtags and their spread properties in the restricted dataset. *b)* shows our derived classes of hashtags according to the spread metrics.

Spatial impact. To get insights of on cities' impact on another, we use the spatial impact metric from [6]. The hashtag specific spatial impact $I^h_{A\to B}$ of two cities A and B and a hashtag h is defined as a score in the range $[-1, 1]$. A score of 1 means that either all occurrences of that hashtag in city A happened before all occurrences in B, or that there are no occurrences of that hashtag in B at all. The same applies in the reverse case scoring -1. Values around 0 indicate that both cities adopted the hashtag roughly at the same time. In short, this measure describes which city adopted a hashtag earlier, and therefore *may* have influenced the other city. The spatial impact $I_{A\to B}$ is then defined as the average hashtag's spatial impact *for all hashtags* that occur in at least one of the cities.

As an example, we compare the cities Aachen, Hamm, and Overath with the 500 most popular cities. For each of the three cities, we show the spatial impact on every of the 500 most popular cities as a histogram in Figure 6. We chose Aachen as the birthplace of the Jodel network with a large technical university and 250 k inhabitants, Hamm as a medium-sized city without university and 180 k inhabitants, and Overath as a smaller city with 27 k inhabitants. The histograms x-axis denotes the spatial impact, while the y-axis covers the number of other cities in comparison. From the given examples, we observe that Aachen is the most influencing city within this comparison (and also on the whole platform Jodel–*not shown*), with most of its scores being between 0.5 and 1. Hamm is both influenced by cities as well as influencing other cities, whereas Overath is heavily influenced by most other cities (probably also due to a low population and therefore fewer users). By also qualitatively looking into other cities spatial impact histogram, we can only conclude that cities with a higher population impact cities



(a) Aachen (b) Hamm (c) Overath

Figure 6: Histograms of spatial impact from Aachen, Hamm, and Overath to the top 500 locations in the complete dataset. Aachen heavily influences most other cities, Overath is mostly influenced by other cities, and Hamm is both influenced by several cities and influencing other cities.

with a lower population. This finding that large cities influence smaller ones is in line with observations on Twitter [6].

We remark that the spatial impact metric does not normalize by community size and thus comparing communities of unequal size can provide an advantage in this metric to the larger city. Even if the hashtags in the big city never spread to any other city, it would still impact a small city using this measure. Nevertheless, this still supports the findings also shown for Twitter that larger cities usually have a higher impact.

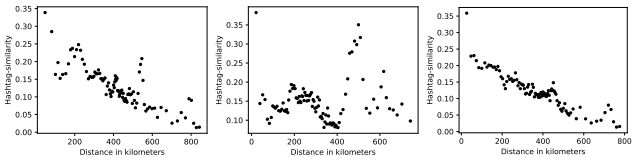
Hashtag similarity. We previously have seen that cities impact each other. To understand the communities hashtags better in comparison, we use the *hashtag similarity* [6] measure of two locations A and B as $\text{sim}(A, B) = |H_A^{50} \cap H_B^{50}|/50$, where H_L^{50} defines the 50 most popular hashtags in location L.

For each location, we calculated the hashtag similarity to all others. Figure 7 shows the results for Aachen, Munich, and Overath in averages for groups of 100 locations. While the x-axis describes the distance to other cities, the y-axis denotes the similarity score. For Aachen and Overath, we observe that closer locations are on average more similar than locations farther away. However, there are several peaks of which the biggest ones represents Berlin[2]. It seems apparent that big cities are connected to each other and share hashtags no matter the distance, which is supported by the example of Munich. Yet, small cities like Overath are less affected. [5] showed similar results for Twitter: W.r.t hashtags, big cities are more similar to each other than to closer, smaller cities.

We verified that this also applies for Jodel considering all hashtags of both cities. The relation we see for Overath of closer cities having more hashtags in common has likewise been shown for Twitter [6]. Our hypothesis is that on Jodel, hashtags *travel* long distances between big cities and then spread across smaller cities within the local neighborhood.

Findings. While the hashtag similarity metric does not directly reflect individual user's contribution to hashtag spreading, it still provides insights into the dis-/similar hashtag usage of communities. Big cities share more popular hashtags and are therefore generally more similar to each other, whereas smaller cities gradually share their most popular hashtags with their local neighborhood. In combination with the spatial influence, this supports our conclusion that hashtags likely spread via the bigger cities into such local neighborhoods.

[2]Within our dataset, Berlin is split into districts and therefore present multiple times.



(a) Aachen (b) Munich (c) Overath

Figure 7: The hashtag similarity of Aachen, Munich, and Overath to cities in a certain distance. Cities closer to each other tend to share more hashtags. However, big cities are similar to each other no matter the distance. Averages of groups of 100 locations.

6 HASHTAG CLASSIFICATION

Within our analysis of hashtags, we have observed that the hashtags can be clustered into different groups (cf. Figure 5a & 5b). We know from literature that there are corresponding types of hashtags on e.g., Twitter. That is, [6] distinguishes between local interest hashtags, regional and event-driven hashtags, and other worldwide memes. We were wondering if and in which way Jodel's locality actually catalyzes other—very local—or prohibits global hashtags. For answering this questions, we create a statistical classifier for determining the hashtag type in three steps: *i)* defining suitable hashtag classes in line with our observations so far, *ii)* manual hashtag classification for providing an answer on a content level, and *iii)* training and validation of statistical models.

6.1 Hashtag Content Categories

Leveraging hints from Section 5.4, manual inspection and expert domain knowledge, we first iteratively defined and verified four different meme classes as follows:

- **Local events:** Often trends originating from a single post (e.g., a funny story) that gained attention in the local community. It is typically very local and short-lived.
- **Local phenomena:** Trend usually related to local persons or buildings. It is typically very local and long-lived.
- **Events:** Short-lived or recurring trend usually related to a real-world happening of larger interest.
- **Other memes:** Memes not included in Jodelstories or Local phenomena.

We labeled the most 450 popular hashtags that had their first occurrence after 1st January 2016 to filter out most of the generic statements. Besides, this makes the classes more balanced, as local trends are much more prominent in this restricted dataset. Due to missing context information or non-fitting classes, we could not classify 49 hashtags. The majority (64 %) of the remaining 401 hashtags were labeled *other meme*, whereas *local phenomenon* (82) represents the second biggest class, Events (35) and Local Event (29) being relatively equal in size.

Having learned that we indeed find trends in terms of hashtags that w.r.t our previous metrics and the manual classification reflect the locality of the Jodel application, we next try to establish the classification methods for them. Thus, we define features that we will use including the presented and analyzed metrics plus some additional temporal and text-based ones in the following section.

Table 2: The features used for the classification.

Feature	Definition
Focus	The focus of the hashtag.
Entropy	The entropy of the hashtag.
Spread	The spread of the hashtag.
Local variation	The local variation of the hashtag. A measure for the regularity of the hashtag's usage.
Hashtags	Average number of hashtags per Jodel.
Comments	Average number of comments per Jodel.
Exclamations	Fraction of Jodels that contain an exclamation mark.
Questions	Fraction of Jodels that contain a question mark.
Temporal focus	The amount of Jodels posted on the peak day of the hashtag divided by the total number of uses.
Temporal entropy	Similar to spatial entropy where different days are considered. Gives a number for the "randomness" of the distribution.
Temporal spread	Similar to spatial spread of the avg distance [days] from the weighted midpoint of all occurrences of the hashtag.
Peak increase	Compares post volume of seven days before the peak with the height of the peak. Is a measure for how "sudden" the peak occurred. A low value indicates a sudden increase in popularity.
Peak decline	Seven days after the peak divided by the height of the peak. Describes how fast interest declined after the peak day. A low value means the interest disappeared suddenly.
User diversity	Number of unique users of the hashtag divided by its total use.

6.2 Features

Our aim is to create a statistical classifier for determining the hashtag type. For our classification approach, we used the features listed in Table 2. This list includes all spatial and temporal metrics that have been discussed before. Besides simple features like hashtag and comment counts, we further added temporal metrics of *peak increase* being defined as the number of posts in seven days prior to the peak divided by the number of posts on the peak day—and *peak decline* alike, but after the peak. These features, therefore, describe how suddenly a trend occurred and disappeared.

6.3 Classifiers and Results

Classifiers. We have applied different statistical methods to our classification problem: k-nearest neighbors, Classification & Regression Trees, Naive Bayes, Logistic Regressen, LDA and ZeroR as a baseline. We used 10-fold cross-validation on our manually classified hashtag dataset to verify the results of each classifier. All classifiers outperform the baseline ZeroR-classifier. While all approaches perform well (detailed results omitted), LDA resulted in a good compromise of the smallest average ± standard deviation. Therefore, we only present the results of the LDA classifier in Table 3. We observe that *events* have the lowest precision value with 0.66. However, this is still a good result as less than 10% of the hashtags are events. The other results are good as well, especially the local phenomena and memes with high F1 scores.

In this classification, both the spatial and the temporal features provided most benefit as removing them caused in both cases a considerable drop in accuracy of at least 0.1, whereas user diversity had only a very minor influence.

Findings. We have shown that we can predict the class of a hashtag by using its spatial and temporal properties. In conclusion, this confirms our theory that the Jodel platform actually has specific local short-lived and long-lived hashtags that differ to countrywide generic memes and events. While we may extend the classification scheme with more features and could apply advanced machine learning techniques, such as neural networks, this is a first step towards automatically classifying certain countrywide/gloabl and in opposition local trends on Jodel—either being short- or long-lived according to our defined classes.

Table 3: Precision, recall and f1-score using the LDA classifier. Averages of 10 runs with different dataset-splits.

	Precision	Recall	F1-Score
Event	0.66	0.80	0.70
Local event	0.79	0.72	0.74
Local phenomena	0.87	0.95	0.91
Other memes	0.97	0.93	0.94

7 CONCLUSIONS

Within this paper, we study the hashtag propagation through the lens of a platform operator by having the unique opportunity to analyze data from Germany (2014 to 2017) provided by Jodel. With this longitudinal data set, we studied the key design pattern of being location-based and its influence on hashtag usage and spreading in comparison to the global counterpart Twitter. We applied established metrics designed to capture the spatial focus and spread of Twitter hashtags [2] to Jodel and extend them with a temporal dimension covering the diffusion of hashtags in time. While we find significant qualitative differences to Twitter of hashtags generally being less focused on Jodel and thus having a higher entropy, the spatial spread also deviates from Twitter. Yet, we find evidence for local hashtags that are a potential result of Jodel's design.

Further, we identify similarities in hashtag usage between nearby and larger cities and present case studies of their spacial impact supporting this finding. By correlating spatial and temporal metrics, we identify four different hashtag classes distinguished by their spatial and temporal extent. Informed by manual labeling of 450 most frequently used hashtags, we created an automatic classification scheme using machine learning models with great success.

While we focused on the empirical birds-eye view on the hashtag usage, it will be interesting trying to apply epidemic modeling approaches. Further, individual user behavior and possible groups w.r.t their spreading influence will provide deeper insights—especially in the sense of Jodel's design choice of being location-based.

Acknowledgement

This work has been funded by the Execellence Initiative of the German federal and state governments.

REFERENCES

[1] Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond Trending Topics: Real-World Event Identification on Twitter *(ICWSM)*.
[2] John Cannarella and Joshua A. Spechler. 2014. Epidemiological modeling of online social network dynamics. *CoRR*.
[3] S. Chandra, L. Khan, and F. B. Muhaya. 2011. Estimating Twitter User Location Using Social Interactions–A Content Based Approach *(SocialCom/PASSAT)*.
[4] P. Alex Dow, Lada A. Adamic, and Adrien Friggeri. 2013. The Anatomy of Large Facebook Cascades. In *ICWSM*.
[5] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. 2013. Traveling Trends: Social Butterflies or Frequent Fliers? *COSN*.
[6] Krishna Y Kamath, James Caverlee, Kyumin Lee, and et al. 2013. Spatio-Temporal Dynamics of Online Memes: A Study of Geo-Tagged Tweets *(WWW)*.
[7] Dimitrios Kotsakos, Panos Sakkos, Ioannis Katakis, and Dimitrios Gunopulos. 2014. #Tag: Meme or Event? *(ASONAM)*.
[8] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the Dynamics of the News Cycle *(KDD)*.

[9] R. Li, K. H. Lei, R. Khadiwala, and K. C. Chang. 2012. TEDAS: A Twitter-based Event Detection and Analysis System *(ICDE)*.
[10] Jonathan Mahler. 2015. Who Spewed That Abuse? Yik Yak Isn't Telling. http://www.nytimes.com/images/2015/03/09/nytfrontpage/scan.pdf. (2015).
[11] Yasuko Matsubara, Yasushi Sakurai, and et al. 2017. Nonlinear Dynamics of Information Diffusion in Social Networks. *ACM Trans. Web*.
[12] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter *(WWW)*.
[13] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors *(WWW)*.
[14] Ceyda Sanlı and Renaud Lambiotte. 2015. Local Variation of Hashtag Spike Trains and Popularity in Twitter. *PLOS ONE* (2015).
[15] Maximilian Walther and Michael Kaisser. 2013. Geo-spatial Event Detection in the Twitter Stream *(Advances in Information Retrieval)*.
[16] Jianshu Weng and Bu-Sung Lee. 2011. Event Detection in Twitter *(ICWSM)*.
[17] Jiyoung Woo and et al. 2016. Epidemic model for inform. diffusion in web forums: experiments in marketing exchange and political dialog. *SpringerPlus* (2016).
[18] Weiai Wayne Xu, Ji Young Park, Ji Young Kim, and et al. 2016. Networked Cultural Diffusion and Creation on YouTube: An Analysis of YouTube Memes. (2016).
[19] Qiang Yan, Lianren Wu, and et al. 2013. Information Propagation in Online Social Network Based on Human Dynamics *(Abstract and Applied Analysis)*.
[20] Savvas Zannettou and et al. 2017. The web centipede: understand. how web communit. infl. each other through the lens of mainstream and alt. news sources *(IMC)*.

MDML'2019 Chairs' Welcome

It is our great pleasure to welcome you to *MDML'2019* at *TheWebConf*. This half-day workshop aims to bring together researchers and practitioners from two research domains - mechanism design and machine learning - whose fields interact billions of times per day in practice but that are still, for the most part, keeping separate in the academic arena. The workshop's goals are to motivate, promote and disseminate interdisciplinary research combining these fields.

We set off to tackle topics at the intersection of the two fields, including the increasing use of data and machine learning in designing mechanisms in broad contexts (e.g., mechanisms based on sampled data or the use of ML techniques in designing new mechanisms), as well as the potential use of mechanism design techniques in machine learning (e.g., learning in a strategic setting). Of special interest are repeated online auctions, where machine learning is an essential component of mechanism design from both the buyers' and sellers' points of view.

The workshop consists of four invited and two contributed talks. Of the six talks, three are from academia and three are industrial talks. This is a testament to the workshop's theme being an active area of theoretical research, as well as having real-life large-scale applications with massive implications.

Our agenda is as follows:

2:00pm	Chairs – welcome remarks
2:05pm	Invited talk by Prof. Tim Roughgarden, Stanford University: "The Surprising Power of Reserve Prices"
2:35pm	Contributed talk by Thomas Nedelec, Noureddine El Karoui and Vianney Perchet, Criteo: "Learning to bid in revenue maximizing auctions"
3:00pm	Invited talk by David Pennock, Microsoft Research: "Smarter Markets: Bringing Intelligence into the Exchange"
3:30pm	--Coffee Break--
4:00pm	Invited talk by Suju Rajan and Noureddine El Karoui, Criteo: "Auction theory from the bidder standpoint"
4:30pm	Contributed talk by Yash Kanoria, Columbia University and Hamid Nazerzadeh, University of Southern California: "Dynamic Reserve Prices for Repeated Auctions: Learning from Bids"
4:55pm	Invited talk by Prof. David Parkes, Harvard University: "Optimal Economic Design through Deep Learning"
5:25pm	Chairs – closing remarks

We thank our Program Committee members for their dedicated and timely input, as well as TheWebConf's Workshop & General Chairs for their support in organizing this workshop.

Aranyak Mehta
MDML'2019 Co-Chair
Google, USA

Ronny Lempel
MDML'2019 Co-Chair
Outbrain, Israel

MDML'2019 Program Committee

Michal Aharon (Oath, Israel)
Omer Ben Porat (Technion, Israel)
Ruggiero Cavallo (Oath, USA)
Olivier Chapelle (Google, USA)
Iftah Gamzu (Amazon, Israel)
Yishay Mansour (Tel-Aviv University , Israel)
Muthu Muthukrishnan (Amazon, USA)
Hamid Nazerzadeh (USC and Uber, USA)
Amir Ronen (SparkBeyond, Israel)
Danny Rosenstein (Outbrain, Israel)
Erik Sodomka (Facebook, USA)
Sergei Vassilvitskii (Google, USA)

Smarter Markets: Bringing Intelligence into the Exchange

David M. Pennock
Microsoft Research
New York City, NY, USA, dpennock@microsoft.com

ABSTRACT

Billions of dollars in financial securities exchange hands every day in independent continuous double auctions. Although the auctions are automated, fast, open 24-7, and have worldwide scope and massive scale, the underlying auction rules have not changed much for over 100 years. Advertisement auctions, on the other hand, have rapidly evolved, incorporating optimization and machine learning directly into their allocation rules. The downside is a less-transparent auction, but the upsides for efficiency and expressiveness are tremendous. The trend toward smarter markets will expand into finance and well beyond, pervading how markets are designed. I will discuss markets that optimize and learn, using prediction markets and advertising markets as key examples.

CCS CONCEPTS

• **Applied computing~Online auctions** • **Applied computing~Economics** • Applied computing~Forecasting • Mathematics of computing~Probability and statistics • Computing methodologies~Artificial intelligence • Computing methodologies~Machine learning

KEYWORDS

Smart markets, prediction markets, advertisement auctions, financial markets, continuous double auction, combinatorial prediction markets, expressive markets, optimization and machine learning in auctions and markets

ACM Reference format:

David M. Pennock. 2019. Smarter Markets: Bringing Intelligence into the Exchange. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316530

1 SPEAKER BIO

David Pennock is a Principal Researcher at Microsoft Research New York City. His largest contributions are novel prediction markets and wagering mechanisms: financial markets harnessed to elicit probabilistic information from a crowd. He has over 70 publications cited 14,000 times and an h-index of 50. He has over twenty patent applications, over twenty press mentions, and has given more than fifty talks. His Ph.D. is in artificial intelligence and he has been an intellectual and organizational leader in the economics-and-computation subfield of AI for two decades. He co-founded two research areas, three workshops, and an ACM journal, and was a founding member of three corporate basic-research labs. He served as Assistant Managing Director of MSR NYC for six years. He was Chair of ACM SIGecom, Program co-Chair of ACM EC, and is co-Editor-in-Chief of ACM TEAC. In addition to his primary research area, he has published work in machine learning, theory, information retrieval (including a Test of Time Award honorable mention in SIGIR), web science, sponsored search, Bayesian networks, constraint satisfaction, and recommender systems. He led the development of several popular online market games and blogged for Yahoo News. In 2005, he was named to MIT Technology Review's list of 35 "top technology innovators under age 35" having the potential to profoundly impact the world.

2 ACKNOWLEDGMENTS

This talk represents joint work with many fantastic colleagues that I have had the pleasure and fortune to work with.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316530

Auction Theory from the Bidder Standpoint

Suju Rajan
Criteo AI Lab
s.rajan@criteo.com

Noureddine El Karoui
Criteo AI Lab and UC, Berkeley
nkaroui@berkeley.edu

ABSTRACT

Much of classical auction theory has been developed from the standpoint of the seller, trying to understand how to optimize auctions to maximize seller revenue for instance. This is still a source of very active current research.

Billions of auctions are now run on the Internet everyday between the same sellers and bidders and this creates a need to better understand auctions from the bidders' perspective.

In this talk we will present some recent results on this question, showing for instance that auctions that are reputed to be truthful are not truthful anymore when the seller optimizes the auction format based on bidders' past bids, provide explicit and simple to implement shading strategies that improve bidders' utility (on and off equilibrium) and are robust to various forms of estimation error and mechanism changes. We will also discuss various equilibrium questions.

We take a mostly functional analytic point of view on these problems. If time permits, we will discuss ongoing work on a machine-learning-based perspective.

Joint work with Thomas Nedelec, Marc Abeille, Clément Calauzènes, Benjamin Heymann and Vianney Perchet while doing research at Criteo.

CCS CONCEPTS

• **Theory of computation → Algorithmic game theory and mechanism design**; **Computational pricing and auctions**; • **Theory of computation Computational advertising theory**;

KEYWORDS

Auction theory; bidder standpoint; truthful auctions; virtual value

ACM Reference Format:
Suju Rajan and Noureddine El Karoui. 2019. Auction Theory from the Bidder Standpoint. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316528

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316528

Incentive-Compatible Learning of Reserve Prices for Repeated Auctions

Yash Kanoria
ykanoria@columbia.edu
Graduate School of Business
Columbia University

Hamid Nazerzadeh
hamidnz@marshall.usc.edu
Marshall School of Business
University of Southern California

ABSTRACT

Motivated by online advertising market, we consider a seller who repeatedly sells ex ante identical items via the second-price auction. Buyers' valuations for each item are drawn i.i.d. from a distribution F that is unknown to the seller. We find that if the seller attempts to dynamically update a common reserve price based on the bidding history, this creates an incentive for buyers to shade their bids, which can hurt revenue. When there is more than one buyer, incentive compatibility can be restored by using *personalized* reserve prices, where the personal reserve price for each buyer is set using the historical bids of *other* buyers. In addition, we use a *lazy* allocation rule, so that buyers do not benefit from raising the prices of their competitors. Such a mechanism asymptotically achieves the expected revenue obtained under the static Myerson optimal auction for F. Further, if valuation distributions differ across bidders, the loss relative to the Myerson benchmark is only quadratic in the size of such differences. We extend our results to a contextual setting where the valuations of the buyers depend on observed features of the items.

ACM Reference Format:
Yash Kanoria and Hamid Nazerzadeh. 2019. Incentive-Compatible Learning of Reserve Prices for Repeated Auctions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316526

Advertising is the main component of the monetization strategy of most Internet companies. A large fraction of online advertisements are sold via auctions where advertisers bid in real time for a chance to show their ads to users. Examples of such auction platforms, called advertisement exchanges (McAfee 2011, Muthukrishnan 2009), include Google's Doubleclick (AdX), Facebook, AppNexus, and OpenX.

The second-price auction is the dominant mechanism used by advertisement exchanges. Among the reasons for its prevalence are the simplicity of the second-price auction and the fact that it incentivizes advertisers to be truthful. The second-price auction can maximize the social welfare (i.e., the value created in the system) by allocating the item to the highest bidder.

To maximize the revenue earned in a second-price auction, the auctioneer can set a reserve price and not make any allocations when the bids are low. In fact, under symmetry and regularity assumptions, the second-price auction with an appropriately chosen reserve price is optimal and maximizes revenue among all selling mechanisms Myerson 1981, Riley and Samuelson 1981.

However, to set the reserve price effectively, the auctioneer requires information about the distribution of the valuations of the bidders. To understand the effects of changing reserve prices based on previous bids, we study a setting where the auctioneer sells impressions (advertisement space) via repeated second-price auctions. We demonstrate that the long-term incentives of advertisers play an important role in the performance of these repeated auctions.

We show that natural mechanisms that set a common reserve price using the history of the bids may create substantial incentives for the buyers to shade their bids. On the other hand, we propose an *incentive-compatible* mechanism that sets a personalized reserve price for each agent based on the bids of *other* agents.[1] Our mechanism allocates the item to the highest bidder if his bid exceeds his personal reserve price. If the item is allocated, the price is equal to the maximum of the second-highest bid and the personal reserve of the winner. This structure corresponds to mechanisms used in practice, as described in Paes Leme et al. 2016. By appropriately choosing the function that maps historical bids of others to a personal reserve price, we show that the expected revenue per round is asymptotically as large as that under the static Myerson optimal auction. Here we build on prior work that investigates how samples from a distribution can be used to set a near-optimal reserve price, cf., Dhangwatnotai et al. 2015. Importantly for incentives, our mechanism is "lazy" (see more on this in the section on related work below), in that it allocates the item only if the highest bidder clears her personal reserve; it is not enough for *some* bidder to clear her personal reserve. An "eager" variant would allocate the item to the highest bidder among those who clear their reserve; this would create an incentive for agents to overbid so as to increase the personal reserves of other agents in future, thereby increasing their likelihood of being eliminated.

A natural concern with such an approach is that if agents' valuation distributions differ from each other, it may lead to a lower personal reserve price for agents with a higher valuation distribution, and vice versa, thereby hurting revenue. We show that this issue is not significant when differences in valuation distributions are not too large. In particular, we show that the loss relative to the Myerson benchmark is only quadratic in the size of such differences, and supplement this theoretical result with numerical examples.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316526

[1] In case of unlimited supply the incentive compatibility directly follows if the price of each buyer only depends on the other buyers, cf., Balcan et al. 2018. With limited supply, obtaining incentive compatibility is more challenging because of "competition" among buyers.

We extend our results to a contextual setting with heterogeneous items that are represented by a feature vector of covariates. The valuations of the buyers are linear in the feature vectors (with a-priori unknown coefficients) plus an idiosyncratic private component. We present a learning algorithm that determines the reserve price for each buyer using an ordinary least squares estimator for the feature coefficients. We show that the loss of revenue is sub-linear in the number of samples (i.e., the number of previous auctions).

We also propose an essentially optimal mechanism that can extract (almost all of) the surplus of the agents. The basic idea is that using the bids of other agents, the seller can construct an estimate of the valuation distribution and hence of the expected utility per round of each agent when individual items are allocated using second-price auctions. Based on this estimate, the mechanism charges a surplus-extracting up-front payment at the beginning of each round. Since agents can influence the up-front payments of other agents, they may have an incentive to overbid to eliminate competing agents from future auctions. We propose a solution that asymptotically removes the incentive for agents to deviate from truthfulness: the mechanism simulates[2] agents who choose not to pay the entrance fee. We show that under our mechanism, truthfulness constitutes an approximate equilibrium.

In each of the mechanisms we propose, we overcome incentive issues using the same two key ideas: (i) we individually choose a pricing rule for each agent, based only on the bids of *other* agents, and (ii) we prevent an agent from benefitting if other agents are prevented from participating by raising the prices they face. In a setting where agents' valuation distributions are identical (or similar to each other), this approach enables the seller to obtain as much revenue as if he knew the valuation distribution F, while maintaining incentive compatibility. We believe that these design principles should be broadly applicable to overcome the lack of knowledge of F when there is competition between strategic agents/buyers.

Finally, we would like to highlight two of our technical contributions that would of broad interest to the mechanism design community: (i) our results on heterogeneous distribution of bidders extends to second price auction with common reserve and other related auctions. In this way, they complement the worst case guarantee (factor 2) of Hartline and Roughgarden 2009. More specifically, as a bi-product of the analysis leading to our main Theorem, we obtain that using a second price auction in the asymmetric valuations case, the seller can obtain expected revenue within $O(\delta^2)$ of the optimal, where δ is the maximum "distance" between valuation distributions. (ii) As part of our analysis on the revenue loss of our mechanism, we prove a lower bound of $\Omega(1/(\#\text{samples})^{2/3})$ up to logarithmic factors on the monopolist's revenue loss when i.i.d. samples are used to learn the optimal reserve price. To the extent of our knowledge, our work provides the first lower-bound for this problem, where the best known upper-bound is of $O(1/(\#\text{samples})^{1/2})$, cf., Dhangwatnotai et al. 2015.

[2]We note that Gorokh et al. 2017, which is an independent and concurrent work to ours, present a similar simulation approach for mechanism design.

REFERENCES

Maria-Florina Balcan, Avrim Blum, Jason D Hartline, and Yishay Mansour. 2008. Reducing mechanism design to algorithm design via machine learning. *J. Comput. System Sci.* 74, 8 (2008), 1245–1270.

Peerapong Dhangwatnotai, Tim Roughgarden, and Qiqi Yan. 2015. Revenue maximization with a single sample. *Games and Economic Behavior* 91 (2015), 318–333.

Artur Gorokh, Siddhartha Banerjee, and Krishnamurthy Iyer. 2017. From Monetary to Non-Monetary Mechanism Design via Artificial Currencies. (2017).

Jason D. Hartline and Tim Roughgarden. 2009. Simple versus Optimal Mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*. 225–234.

Preston McAfee. 2011. The Design of Advertising Exchanges. *Review of Industrial Organization* 39, 3 (2011), 169–185.

S. Muthukrishnan. 2009. Ad Exchanges: Research Issues. In *Internet and Network Economics, 5th International Workshop (WINE)*. 1–12.

Roger B. Myerson. 1981. Optimal Auction Design. *Mathematics of Operations Research* 6, 1 (1981), 58–73.

Renato Paes Leme, Martin Pál, and Sergei Vassilvitskii. 2016. A Field Guide to Personalized Reserve Prices. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1093–1102.

John G. Riley and William F. Samuelson. 1981. Optimal Auctions. *American Economic Review* 71, 3 (1981), 381–392.

Learning to Bid in Revenue Maximizing Auction

Extended Abstract

Thomas Nedelec
Criteo AI Lab
ENS Paris Saclay
thomas.nedelec@polytechnique.edu

Noureddine El Karoui
UC, Berkeley
Criteo AI Lab
nkaroui@berkeley.edu

Vianney Perchet
ENS Paris Saclay
Criteo AI Lab
vianney.perchet@normalesup.org

ABSTRACT

We consider the problem of the optimization of bidding strategies in prior-dependent revenue-maximizing auctions, when the seller fixes the reserve prices based on the bid distributions. Our study is done in the setting where one bidder is strategic. Using a variational approach, we study the complexity of the original objective and we introduce a relaxation of the objective functional in order to use gradient descent methods. Our approach is simple, general and can be applied to various value distributions and revenue-maximizing mechanisms. The new strategies we derive yield massive uplifts compared to the traditional truthfully bidding strategy.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Online auctions**; • **Mathematics of computing** → Nonconvex optimization.

ACM Reference Format:
Thomas Nedelec, Noureddine El Karoui, and Vianney Perchet. 2019. Learning to Bid in Revenue Maximizing Auction: Extended Abstract. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316527

INTRODUCTION

Modern marketplaces like Uber, Amazon or Ebay enable sellers to fine-tune their selling mechanism by reusing their large number of past interactions with consumers. In the online advertising or the electricity markets, billions of auctions are occurring everyday between the same bidders and sellers. Based on the data gathered, different approaches learn complex mechanisms maximizing the seller revenue [5, 9, 18, 19].

Most of the literature has focused on the auctioneer side [14]. Algorithms focused on the bidder's standpoint to enable them to be strategic against any smart data-driven selling mechanisms are lacking. These algorithms should ideally strengthen the balance of power driving the relationship between buyers and sellers. Our main objective is to exhibit simple robust algorithmic procedures that take advantage of various data-dependent revenue-maximizing mechanisms. This represents a big step forward in understanding possible strategic behaviors in revenue maximizing auctions. This is a new argument supporting the Wilson doctrine [22] claiming that data-dependent revenue maximizing algorithms are not robust to strategic bidders.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316527

Framework

In the early stage of the market design literature (see, e.g., Myerson [17]), a typical underlying assumption is that the bidders' value distributions were commonly known to the seller and other bidders. This can be justified if different group of bidders with the same value distribution are interacting successively with one seller. In the aforementioned modern applications, the same bidders have billions of interactions everyday with the seller. Even if the latter does not know the value distribution beforehand, it might use in many cases the past bid distributions as proxies of value distribution.

Several mechanisms based on the value distribution of bidders have already been introduced. We will focus on the lazy second price auction with personalized reserve price [19], the Myerson auction [17], the eager version of the second price auction and the boosted second price auction [9]. When repeating these auctions (every day, or every milli-second, depending on the context) and if the bidder is *myopic*, i.e optimizing per stage and not long-term revenue, it is optimal to bid truthfully at each auction. So with myopic bidders, bids and values have the same distribution and the seller can design optimally the mechanism based on the former.

Non-myopic bidders optimize their *long-term expected utility* taking into account that their current strategy will imply a certain mechanism (for instance a specific reserve price) in the future. More precisely, we will consider the following steady state analysis. Assume the valuations of a bidder $v_i \in \mathbb{R}$ are drawn from a specific distribution F_i; a bidding strategy is a mapping β_i from $\mathbb{R}$ into $\mathbb{R}$ that indicates the actual bid $B_i = \beta_i(v_i)$ when the value is v_i. As a consequence, the distribution of bids F_{B_i} is the push-forward of F_i by β_i. In the steady state, the seller uses the distributions of bids F_{B_i} to choose a specific auction mechanism $\mathcal{M}(F_{B_i})$ among a given class of mechanisms $\mathcal{M}$. The objective of *a long-term strategic bidder* is to find her strategy β_i that maximizes her expected utility when $v_i \sim F_i$, she bids $\beta_i(v_i)$ and the induced mechanism is $\mathcal{M}(F_{B_i})$. This steady-state objective is particularly relevant in modern applications as most of the data-driven selling mechanisms are using large batches of bids as examples to update their mechanism.

In terms of game theory, these interactions are a game between the seller - whose strategy is to pick a mechanism design that maps bid distributions to reserve prices - and the bidders - who chose bidding strategies. Our overarching objective is to derive the best-response, for a given bidder i, to the strategy of the seller (i.e., a given mechanism) and the strategies of the other bidders (i.e., their bid distributions).

Contributions

Our main contributions are the following. We first introduce the optimization problem that strategic bidders are facing when the seller is optimizing personalized reserve prices based on their bid distributions. A straightforward optimization can fail because the objective is discontinuous as a function of the bidding strategy.

To circumvent this issue, we introduce a new relaxation of the problem which is stable to local perturbations of the objective function and computationally tractable and efficient. We numerically optimize this new objective through a simple neural network and get very significant improvements in bidder utility compared to truthful bidding. We also provide a theoretical analysis of *thresholded strategies* and show their (local) optimality as improvements of bidding strategies with non-zero reserve value.

For the Myerson auction, the strategies learned by the model can be independently proved to be optimal. We apply the approach to other auction settings such as boosted second price or eager second price with monopoly price. We report massive uplifts compared to the traditional truthful strategy advocated in all these settings. Our simple approach can be plugged in any modern bidding algorithms learning distribution of the highest bid of the competition and we test it on other classes of mechanism without any known closed form optimal bidding strategies. We finally provide the code in PyTorch that has been used to run the different experiments. This approach opens avenues of research for designing good bidding strategies in many data-driven revenue-maximizing auctions.

Related work

Starting with the seminal work of Myerson [17], a rich line of work indicates the type of auctions that is revenue-maximizing for the seller. In the case of symmetric bidders [17], one revenue maximizing auction is a second price auction with a reserve price equal to the monopoly price, i.e, the price r that maximizes $r(1 - F(r))$. However, in most applications, the symmetric assumption is not satisfied [9]. In the asymmetric case, the Myerson auction is optimal [17] but is difficult to implement in practice [16]. In this case, a second price auction with a well-chosen vector of reserve prices guarantees at least one-half of the optimal revenue [10].

In modern markets, some bidders are myopic simply because truthful bidding is a simple strategy to implement. Receiving truthful bid enables sellers to design various revenue maximizing auctions. [5] has therefore been interested in the automatic mechanism design that fine tunes mechanism based on some examples of bids. This work was extended recently in [7] with the use of deep learning. In [13, 18, 19], it is shown specifically how to learn the optimal reserve prices in the lazy second price auction. This practice was theoretically addressed by [4, 6, 11] looking at the sample complexity of a large class of auctions assuming an oracle offering iid examples of the value distribution.

However, it is quite intuitive that non-myopic bidders should not bid truthfully. Robustness to strategic bidders has been studied in [3, 8, 12]. A potential limitation of this type of approach is that it is either assumed that all bidders have the same value distribution (or up to ε for some specific metric on distributions) or that there is a very large number of bidders and a global mechanism designed so that any of them has no incentive to bid untruthfully. In [2], an involved mechanism was designed that keeps the incentive compatibility property even if the seller is learning on former bids.

None of these papers have exhibited optimal strategies that can be used when the seller is optimizing her mechanism based on past bids. This strategic behavior has been studied for posted price with one bidder and one seller [15]. An independent line of work has focused on learning to bid when the value is not known to the bidders [21, 23]. Some Bayes-Nash equilibria corresponding to games where bidders can choose their bid distribution were designed [1, 20] with some derivations of seller revenue and bidders utility at these equilibria. However, no strategies corresponding to these equilibria were provided in the general case. Our new optimization pipeline is very general and enables bidders to learn good bidding strategies in multiple settings and for any value distribution.

REFERENCES

[1] Marc Abeille, Clément Calauzènes, Noureddine El Karoui, Thomas Nedelec, and Vianney Perchet. 2018. Explicit shading strategies for repeated truthful auctions. *arXiv preprint arXiv:1805.00256*.

[2] Itai Ashlagi, Constantinos Daskalakis, and Nima Haghpanah. 2016. Sequential mechanisms with ex-post participation guarantees. In *Proceedings of EC 2016*.

[3] Santiago R Balseiro, Vahab S Mirrokni, and Renato Paes Leme. 2017. Dynamic mechanisms with martingale utilities. In *Management Science*.

[4] Richard Cole and Tim Roughgarden. 2014. The sample complexity of revenue maximization. In *Proceedings of Theory of computing*.

[5] Vincent Conitzer and Tuomas Sandholm. 2002. Complexity of mechanism design. In *Proceedings of UAI 2002*. Morgan Kaufmann Publishers Inc., 103–110.

[6] Nikhil R Devanur, Zhiyi Huang, and Christos-Alexandros Psomas. 2016. The sample complexity of auctions with side information. In *Proceedings of Theory of Computing*.

[7] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, and David C Parkes. 2017. Optimal auctions through deep learning. *arXiv preprint arXiv:1706.03459* (2017).

[8] Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, and Song Zuo. 2018. Incentive-aware learning for large markets. In *Proceedings of the 27th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee*.

[9] N. Golrezaei, M. Lin, V. Mirrokni, and H. Nazerzadeh. 2017. Boosted Second-price Auctions for Heterogeneous Bidders. In *Management Science*.

[10] Jason D Hartline and Tim Roughgarden. 2009. Simple versus optimal mechanisms. In *Proceedings of EC 2009*.

[11] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. 2018. Making the most of your samples. *SIAM J. Comput.*

[12] Yash Kanoria and Hamid Nazerzadeh. 2014. Dynamic Reserve Prices for Repeated Auctions: Learning from Bids. In *Proceedings of WINE 2014*.

[13] Andres M Medina and Mehryar Mohri. 2014. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *Proceedings of ICML 2014*.

[14] Paul R Milgrom and Steven Tadelis. 2018. *How Artificial Intelligence and Machine Learning Can Impact Market Design.* Technical Report. National Bureau of Economic Research.

[15] Mehryar Mohri and Andres Munoz. 2015. Revenue optimization against strategic buyers. In *Proceedings of NIPS 2015*.

[16] Jamie H Morgenstern and Tim Roughgarden. 2015. On the pseudo-dimension of nearly optimal auctions. In *Proceedings of NIPS 2015*.

[17] R. B. Myerson. 1981. Optimal Auction Design. *Math. Oper. Res.* 6, 1.

[18] M. Ostrovsky and M. Schwarz. 2011. Reserve prices in internet advertising auctions: A field experiment. In *Proceedings of EC 2011*.

[19] Renato Paes Leme, Martin Pal, and Sergei Vassilvitskii. 2016. A field guide to personalized reserve prices. In *Proceedings of WWW 2016*.

[20] Pingzhong Tang and Yulong Zeng. 2018. The price of prior dependence in auctions. In *Proceedings of EC 2018*.

[21] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. 2016. Online learning in repeated auctions. In *Conference on Learning Theory*. 1562–1583.

[22] R. Wilson. 1987. Game-theoretic analyses of trading processes. In *Advances in Economic Theory*.

[23] Chara Podimata Zhe Feng and Vasilis Syrgkanis. 2018. Learning to bid without knowing your value. In *Proceedings of EC 2018*.

MEPDaW Chairs' Welcome

It is our great pleasure to welcome you to the 5th workshop on Managing the Evolution and Preservation of the Data Web, *WWW 2019 MEPDaW*.

The workshop aims at addressing challenges and issues on managing Knowledge Graph evolution and preservation by providing a forum for researchers and practitioners to discuss, exchange and disseminate their ideas and work, to network and cross-fertilise new ideas.

We received 6 submissions from all around the world covering a broad range of topics. We evaluated them regarding relevance, quality, and novelty, selecting 5 papers. We invited 2 extended abstracts from the recent Special Issue of the Journal of Web Semantics on Managing the Evolution and Preservation of the Data Web (January 2019). In addition, Laure Berti-Équille will give an inspiring talk, and we will close the workshop with a plenary discussion. Thus, MEPDaW includes the following talks:

- **ML-Based Knowledge Graph Curation: Current Solutions and Challenges (Keynote)**
  Laure Berti-Équille (IRD / Aix-Marseille Université)
- **RDF Graph Stores as Convergent Datatype Instances**
  James Anderson (Datagraph GmbH)
- **Modelling the Provenance of Linked Data Interlinks for the Library Domain**
  Lucy McKenna (ADAPT, Trinity College Dublin); Christophe Debruyne (ADAPT, Trinity College Dublin); Declan O'Sullivan (ADAPT, Trinity College Dublin)
- **Conflict Detection, Avoidance, and Resolution in a Non-Linear RDF Version Control System: The Quit Editor Interface Concurrency Control**
  Natanael Arndt (AKSW Leipzig Univ.); Norman Radtke (AKSW Leipzig Univ.)
- **Open traffic lights: a strategy for publishing and preserving traffic lights data**
  Brecht Van de Vyvere (Ghent Univ.); Pieter Colpaert (Ghent Univ.); Ruben Verborgh (Ghent Univ.)
- **Summarizing Entity Temporal Evolution in Knowledge Graphs**
  Mayesha Tasnim (RWTH Aachen); Diego Collarana (Fraunhofer IAIS); Damien Graux (Fraunhofer IAIS); Fabrizio Orlandi (ADAPT, Trinity College Dublin); Maria-Esther Vidal (Univ. Simon Bolivar and TIB)
- **Decentralized Collaborative Knowledge Management using Git (Extended Abstract)**
  Natanael Arndt (AKSW Leipzig Univ.); Patrick Naumann (AKSW Leipzig Univ.); Norman Radtke (AKSW Leipzig Univ.); Michael Martin (AKSW Leipzig Univ.); Edgard Marx (AKSW Leipzig Univ.)
- **TISCO: Temporal Scoping of Facts (Extended Abstract)**
  Anisa Rula (Univ. Milano-Bicocca); Matteo Palmonari (Univ. Milano-Bicocca); Simone Rubinacci (Univ. Oxford); Axel-Cyrille Ngonga Ngomo (Paderborn Univ.); Jens Lehmann (Univ. Bonn); Andrea Maurino (Univ. Milano-Bicocca); Diego Esteves (Univ. Bonn)

We hope that you will find the workshop program interesting. Fostering active usage of evolving knowledge graphs requires further research advances on topics such as storage, synchronisation, change representation and querying. Solutions to these problems correspond to the main subjects of interest of the workshop.

Javier D. Fernández
MEPDaW Co-Chair

Jeremy Debattista
MEPDaW Co-Chair

Fabrizio Orlandi
MEPDaW Co-Chair

Maria-Esther Vidal
MEPDaW Co-Chair

THE WEB CONFERENCE

MEPDaW Organization

Workshop Chairs:
Javier D. Fernández *(Vienna University of Economics and Business, Austria)*
Jeremy Debattista *(ADAPT Centre, Trinity College Dublin, Ireland)*
Fabrizio Orlandi *(ADAPT Centre, Trinity College Dublin, Ireland)*
Maria-Esther Vidal, *(Universidad Simon Bolivar and Technische Informationsbibliothek (TIB), Germany)*

Advisory Board:
James Anderson *(Dydra, Germany)*
Rob Brennan *(ADAPT Centre, Dublin City University, Ireland)*
Axel Polleres *(Vienna University of Economics and Business, Austria)*
Jürgen Umbrich *(Vienna University of Economics and Business, Austria)*

Program Committee:
Maribel Acosta *(Karlsruhe Institute of Technology (KIT), Germany)*
Natanael Arndt *(AKSW, Leipzig, Germany)*
Judie Attard *(ADAPT Centre, Trinity College Dublin, Ireland)*
Rob Brennan *(ADAPT Centre, Dublin City University, Ireland)*
Ioannis Chrysakis *(FORTH-ICS, Greece)*
Diego Collarana *(Fraunhofer IAIS and University of Bonn, Germany)*
Christophe Debruyne *(ADAPT Centre, Dublin City University, Ireland)*
Kemele Endris *(L3S, Leibniz University of Hannover)*
Valeria Fionda *(University of Calabria, Italy)*
Giorgos Flouris *(FORTH-ICS, Greece)*
Michael Galkin *(Fraunhofer IAIS, Germany)*
Damien Graux *(Fraunhofer IAIS, Germany)*
Luis Ibañez-Gonzalez *(University of Southampton, UK)*
Sabrina Kirrane *(Vienna University of Economics and Business, Austria)*
Steffen Lohmann *(Fraunhofer IAIS, Germany)*
Marios Meimaris *(ATHENA R.C., Greece)*
Declan O'Sullivan *(ADAPT Centre, Trinity College Dublin, Ireland)*
Harshvardhan J. Pandit *(ADAPT Centre, Trinity College Dublin, Ireland)*
George Papastefanatos *(ATHENA R.C., Greece)*
Giuseppe Pirro *(ICARCNR, Italy)*
Axel Polleres *(WU Vienna, Austria)*
Ruben Taelman *(Ghent University, Belgium)*

THE WEB CONFERENCE

ML-Based Knowledge Graph Curation: Current Solutions and Challenges

Laure Berti-Equille [1,2]
[1]ESPACE-DEV/IRD, UMR 228, IRD/UM/UG/UR, Montpellier, France
[2]Aix Marseille Université, Université de Toulon, CNRS, LIS, DIAMS, Marseille, France
laure.berti@ird.fr

ABSTRACT

With the success of machine learning (ML) techniques, ML has already proved a tremendous potential to impact the foundations, algorithms, and models of several data management tasks, such as error detection, data quality assessment, data cleaning, and data integration. In Knowledge Graphs, part of the data preparation and cleaning processes, such as data linking, identity disambiguation, or missing value inference and completion could be automated by making a ML model "learn" and predict the matches routinely with different degrees of supervision. This talk will survey the recent trends of applying machine learning solutions to improve and facilitate Knowledge Graph curation and enrichment, as one of the most critical tasks impacting Web search and query-answering. Finally, the talk will discuss the next research challenges in the convergence of machine learning and management of Knowledge Graph evolution and preservation.

CCS CONCEPTS

• **Information systems → Data cleaning**; **Semantic web description languages**; *Graph-based database models*; • **Computing methodologies → Machine learning approaches**.

KEYWORDS

Knowledge base curation; knowledge graph completion; entity disambiguation

ACM Reference Format:
Laure Berti-Equille [1,2]. 2019. ML-Based Knowledge Graph Curation: Current Solutions and Challenges. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316522

1 INTRODUCTION

Over the past few years, massive amounts of world knowledge have been accumulated in Web knowledge bases, both commercial and openly available such as Freebase, DBpedia, and YAGO, among the most prominent ones. These KBs contain millions of entities (e.g., people, places, or organizations) and millions of related facts expressed in RDF (Resource Description Framework) in the form of triples such as < `head entity, relation, tail entity`> indicating the relation between two entities.

Knowledge graphs are often constructed from semi-structured knowledge as Wikipedia or harvested from the Web with a combination of statistical and NLP methods. However, Web Data used for knowledge base construction is noisy, unreliable, highly imbalanced, heterogeneous, and evolve over time [1]. The result are large-scale knowledge graphs that try to make a good trade-off between completeness and correctness [3].

Not surprisingly, Web knowledge graphs suffer from a wide range of anomalies: they are greatly incomplete (i.e., with missing entities and links), some entities are ambiguous, and identity relations between resources that refer to the same real world entity are erroneous.

Building high-quality knowledge bases critically depends on the data curation technologies, which are rapidly evolving with the recent advances in representation learning and machine learning. These advances are now leveraged for knowledge base completion, refinement, entity linking, and entity disambiguation.

In this talk, we will survey the recent trend of applying machine learning solutions to improve knowledge graph curation tasks and establish new paradigms to sharpen knowledge graph error detection and cleaning, as an attempt to fill the gap in knowledge curation science [4]. We will discuss the advantages and limitations of these techniques and their extensions to address the current challenges, not only in extracting knowledge from both structured and unstructured data, across a large variety of domains, and in multiple languages, but also in maintaining high quality in evolving knowledge repositories.

2 SPEAKER'S BIO

Laure Berti-Équille is a Research Director at IRD, the French research institute for sustainable development, currently leading the research group DIAMS (Data Integration, Analysis, and Management as Services, http://diams.lis-lab.fr at Aix-Marseille University. Before, she was a full professor at Aix-Marseille University (AMU), senior scientist at Qatar Computing Research Institute (Hamad Bin Khalifa University, Qatar), Associate Professor (with tenure) at University of Rennes 1 (France), and visiting researcher at AT&T Labs Research (USA) as a recipient of the prestigious European Marie Curie Outgoing Fellowship. Her interests are at the intersection of large-scale data analytics, and statistical machine learning with a focus on data quality, anomaly detection, and truth discovery, with more than 80 publications and three monographs. She initiated the very first workshop editions on information and data quality in information systems (IQIS 2005) and quality in databases (QDB 2009 and 2016) in conjunction with SIGMOD and VLDB respectively,

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316522

and co-organized the first French workshops on Data and Knowledge Quality in conjunction with EGC (Extraction et Gestion de Connaissances) in 2005, 2006, 2010, and 2011. She has received various grants from the French Agency for National Research (ANR), the French National Research Council (CNRS), and the European Union.

REFERENCES

[1] Berti-Equille L., Scannapieco M. (2016). Quality of Web Data (Chapter). In the 2nd Edition of the book Data Quality: Concepts, Methodologies and Techniques, Springer, 2016

[2] Zaveri A., Maurino A., Berti-Equille L. (2014). Web Data Quality: Current State and New Challenges, Int. J. Semant. Web Inf. Syst.,10(2):1552-6283, IGI Global.

[3] Paulheim H. (2017) Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8 3 489-508.

[4] Paritosh P. (2018). The Missing Science of Knowledge Curation (Improving incentives for large-scale knowledge curation). In Companion of The Web Conference 2018, Lyon, France.

RDF Graph Stores as Convergent Datatypes

James Anderson
Datagraph GmbH
james@dydra.com

ABSTRACT

This report describes a way to represent and operate on an RDF dataset such the it behaves as an instance of a conflict free replicable datatype. In this industry presentation, we describe how we accomplish this for the Dydra RDF graph storage service in a manner compatible with the SPARQL Graph Store HTTP Protocol (GSP). The standard GSP concerns the current store state only. Dydra retains previous store states as active addressable aspects analogous to named graphs in a quad store. It incorporates and addresses arbitrary revisions of target datasets according to `ETag` and `Content-Disposition` specifications in HTTP headers. Appropriate interpretation of these arguments permits to replicate datasets among cooperating participants.

CCS CONCEPTS

• **Information systems → Version management**; **Resource Description Framework (RDF)**; *Cloud based storage*; *Distributed storage*; *Digital libraries and archives.*

KEYWORDS

CvRDT; Graph Store Protocol; revisions; RDF;
temporal data

ACM Reference Format:
James Anderson. 2019. RDF Graph Stores as Convergent Datatypes. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3316517

1 INTRODUCTION

Dydra is an RDF graph storage service. It operates as a cloud service, a local service or an embedded library. It stores and retrieves the contents of RDF datasets through SPARQL[1], GraphQL[14], TPF[2]and LDP[3] interfaces. It provides uniform access through the SPARQL query language to a variety of data sources, from local RDF quad stores, to local and remote SQL[13] stores, and remote REST[4] data servers. Dydra retains previous store states, in addition to the current state, as active addressable aspects of a dataset analogous to named graphs in a quad store by annotating statements with temporal attributes[5] and permits to address these states as static projections through provision of a `REVISION` argument in SPARQL and Graph Store requests or as streams over MQTT[16] through provision of a `REVISION-WINDOW` argument. Through an appropriate interpretation of revision states, in combination with the appropriate state transfer protocol, it is possible to comprehend repositories as components of a distributed store which behaves as an instance of a conflict-free replicated data type and use them as the storage substrate for distributed collaborative document management.

This document describes how Dydra's schematic repository storage model and Graph Store Protocol extensions combine to provide this capability. The next section describes a schematic graph revision model which fulfils the requirements of a conflict-free replicable data type. Section 3 introduces the Graph Store Protocol and describes extended requests which convey annotations necessary to replicate graph revisions. Section 4 proposes an approach to establish semantically correct states.

2 A MODEL FOR GRAPH REVISIONS AS CRDT

Lamport [19] describes the requirements to order events in a distributed system. He intends to afford causality to the process by deriving a total order for the events in order to achieve "correct" operation, where "correct" intends : all participants agree as to one causality, in which any state depends only on states which "precede" it. As one example, where correctness matters, he discusses how to arbitrate access to a shared resource. For this purpose he proposes, "[i]t just involves making sure that each process learns about all other processes' operations." [19, p.151] For a temporal RDF storage service, in Lamport's terms the shared resource, becomes the "$repositoryState@time >= T_i$". It can achieve a form of *eventual correctness*[6] and derive it retrospectively, rather than àpriori, by relying on asynchronous local clocks and convergent replicable data types. A participant may assert states for times after they have been admitted to the process only and assertions take the form of statement additions and removals with annotations to specify their provenance. Based upon these, replicas will converge globally on a state which reflects causality.

[1]The *SPARQL* Protocol and RDF Query Language (SPARQL) is a query language and protocol for RDF[8]-
[2]"*Triple Pattern Fragments* (TPF), a low-cost interface to triples"[30] within the *Linked Data Fragments* framework[29].
[3]"*Linked Data Platform* (LDP) defines a set of rules for HTTP operations on web resources, some based on RDF, to provide an architecture for read-write Linked Data on the web"[9]

[4]The *Representational State Transfer* (REST) style is an abstraction of the architectural elements within a distributed hypermedia system... in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data elements[12]. Dydra supports Web API endpoints which can be expressed in terms of the Hydra vocabulary[20].
[5]Dydra implements a transaction-time temporal RDF store[1][28] with a particular form of what is characterised as a "timestamp-based approach"[10].
[6] The approach relaxes the requirement, that a local clock be adjusted to reflect remote timestamps[19, p.560].

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316517

The treatments of RDF graphs as CRDT follow from Shapiro's *Strong Eventual Consistency* (SEC) model and its application to set and graph data types. Within this model, if an object supports commutative, associative and idempotent operators, it qualifies as a Convergent Replicated Data Type (CvRDT) [26, p.6][27]. Initial work by Weiss[31] and Nedelec[23] applied this approach to collaboratively edited text documents to demonstrate that is is possible to satisfy CRDT ordering requirements by replicating identifiers which are generated locally by incrementally uniquely modifying existing identifiers. Bieniusa [4] and Deftu [7] described more effective implementations of Shapiro's Observed-Removed-(OR)-Set. An RDF dataset is a set of statements which model a graph [6]. The standard HTTP Graph Store access protocols[24] permit to add to and remove statements from graph. For an RDF dataset which follows just the standard abstract semantics, these operations do not fulfil all of the requirements to qualify as a CvRDT. The operations insertion and deletion are idempotent and associative, but they are not commutative. Aslan's C-Set [2], Ibanez's SU-Set [18][17] and Zarzour's B-Set[32] extend the OR-Set and apply it to distributed semantic stores by following the example of local identifier generation and associating some combination of a presence indicator and provenance record with each operand statement in order to satisfy the SEC requirements and allow operations to converge on a shared state.

As a revisioned RDF dataset extends the standard semantics to distinguish versions, it is instead possible to order operations according to the rules for asynchronous clocks and, with that ordering, to render the operations commutative. This is possible, if one ensures that each modification is labeled in such a manner that a means exists for all participants to establish the same subsequent state given the same eventual knowledge of events. This, in turn is possible if the labels are sufficient to place all operations in a space from which all participants can establish the same total order for events in order to produce the same projection for any given location in the space. This is accomplished by establishing revision identifiers which follow Lamport's rules for distributed clocks: each revision is identified by combinations of

$$localTimestamp \times participantId \times (ADD + REMOVE) \quad (1)$$

and ordering them by first comparing the time value, then the mode number and then the operation, for which ADD precedes REMOVE. As the store is revisioned, all states are retained and the behavior is that of a Last-WriterWins(LWW)-element-Set [25, p. 24], but without the tombstone overhead.

3 GRAPH STORE PROTOCOL EXTENSIONS

The core of the Dydra RDF service is a revisioned quad store, which is accessed through the W3C RDF Graph Store (GSP) and SPARQL protocols. As an extension to standard RDF semantics [6], the store implements the versioning pattern, as described above ("timestamp/participant/operation"), In order to permit operations which express the terms described above, the information must be conveyed in Graph Store Protocol requests. This means any extensions must be compatible with the GSP/HTTP protocol at the protocol and the application level.

The standard GSP update methods are

```
PATCH /jhacker/test/propagate HTTP/1.1
Content-Type: multipart/related; boundary=PATCH
Etag: d745f480-2661-11e9-9eb9-010203040506
Content-Disposition: replicate=d3graph

--PATCH
X-HTTP-Method-Override: DELETE
ContentType: application/n-quads

<http://example.org/5f3bc330-2052-11e9-82ae-010203040506>\
 <http://example.org/x>\
  ``963''^^<http://www.w3.org/2001/XMLSchema#integer> .
--PATCH
X-HTTP-Method-Override: POST
ContentType: application/n-quads

<http://example.org/5f3bc330-2052-11e9-82ae-010203040506>\
 <http://example.org/y>\
  ``291''^^<http://www.w3.org/2001/XMLSchema#integer> .
--PATCH--
```

Figure 1: HTTP PATCH request

- **DELETE**: delete the request content from the the graph.
- **POST**: add the request content to the content of the graph.
- **PUT**: replace the content of the graph with the request content.

We extend these to

- use multipart/related content as a sequence of sub-GSP operations in a **PATCH** request to modify store content,
- use an `ETag` header to specify revision identifiers, and
- use a `Content-Disposition` request header to define, subscribe to and route **PATCH** requests.

Where these requests identify no revision, a revision identifier is generated which combines the current participant-time position of the service node with the operation identifier. If a revision identifier is provided, that is asserted as the participant-time position. In order to provide for the case where an operation comprises both additions and deletions, the **PATCH** operation is also supported in a manner which accepts multipart/related request content to perform the composite operation within a single store transaction. Such a request would take the form illustrated in figure 1[7]. It replaces the 'x' property of a node by sequencing a **DELETE** and a **POST** operation and positions the result revision by providing an `ETag`. The `Content-Disposition` identifies an exchange which indexes WebSocket[11] connections to participants over which the **PATCH** requests are sent in order to replicate state.

4 APPLICATION-LEVEL SEMANTIC INTEGRITY

The CRDT approach guarantees eventual consistency between trusted nodes at the RDF graph level. It does not address malicious or untrusted participants at the application level and it does not provide means to address application level semantic constraints, e.g. double spend race conditions. We propose to mitigate malicious operations by recording each parent revision in order to capture provenance chains and suppressing any states asserted by such participants based on the location information encoded in the revision identifier. Where the revisions from a specific location are suppressed, a temporal projection is then computed as if they were

[7]The example depicts N-Quads[5], with the lines are wrapped for presentation. Where structural anonymous nodes[21, p. 425] are required, Turtle[3] can be used to distinguish identifiers with document scope and dynamic extent from those with indefinite extent and a scope which comprises the surface shared among the replicas[15].

absent, their effect is eliminated from any projection and their support is removed from any dependant revision. Correctness is then established by electing the states which derive respectively from the earliest independent parent revision.

REFERENCES

[1] James Anderson and Arto Bendiken. 2016. Transaction-Time Queries in Dydra. In *Joint Proceedings of the 2nd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2016) and the 3rd Workshop on Linked Data Quality (LDQ 2016) co-located with 13th European Semantic Web Conference (ESWC 2016), Heraklion, Crete, Greece, May 30th, 2016. (CEUR Workshop Proceedings)*, Jeremy Debattista, Jürgen Umbrich, Javier D. Fernández, Anisa Rula, Amrapali Zaveri, Magnus Knuth, and Dimitris Kontokostas (Eds.), Vol. 1585. CEUR-WS.org, 11–19. http://ceur-ws.org/Vol-1585/mepdaw2016_paper_02.pdf

[2] Khaled Aslan, Pascal Molli, Hala Skaf-Molli, and Stéphane Weiss. 2011. C-set: a commutative replicated data type for semantic stores. In *RED: Fourth International Workshop on REsource Discovery*.

[3] David Becket, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. 2014. *RDF 1.1 Turtle*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/turtle/

[4] Annette Bieniusa, Marek Zawirski, Nuno Preguiça, Marc Shapiro, Carlos Baquero, Valter Balegas, and Sérgio Duarte. 2012. An optimized conflict-free replicated set. *arXiv preprint arXiv:1210.3368* (2012). https://arxiv.org/pdf/1210.3368.pdf

[5] Gavin Carothers. 2014. *RDF 1.1 N-Quads*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/turtle/

[6] Richard Cyganiak, David Wood, and Markus Lanthaler. 2014. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/rdf11-concepts/

[7] Andrei Deftu and Jan Griebsch. 2013. A scalable conflict-free replicated set data type. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. IEEE, 186–195.

[8] Lee Feigenbaum, Gregory Williams, Kendall Clark, and Torres Elias. 2013. *SPARQL 1.1 Protocol*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/sparql11-protocol/

[9] Lee Feigenbaum, Gregory Williams, Kendall Clark, and Torres Elias. 2015. *Linked Data Platform 1.0*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/ldp/

[10] Javier D. Fernández, Jürgen Umbrich, Axel Polleres, and Magnus Knuth. 2019. Evaluating query and storage strategies for RDF archives. *Semantic Web* 10, 2 (2019), 247–291. https://doi.org/10.3233/SW-180309

[11] Ian Fette and Alexey Melnikov. 2011. The WebSocket Protocol. *RFC* 6455 (2011), 1–71. https://doi.org/10.17487/RFC6455

[12] Roy Thomas Fielding. 2000. *Architectural styles and the design of network-based software architectures. 2000*. Ph.D. Dissertation. University of California, Irvine. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

[13] James R Groff, Paul N Weinberg, et al. 2002. *SQL: the complete reference*. Vol. 2. McGraw-Hill/Osborne.

[14] Olaf Hartig and Jorge Pérez. 2017. An Initial Analysis of Facebook's GraphQL Language. In *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017. (CEUR Workshop Proceedings)*, Juan L. Reutter and Divesh Srivastava (Eds.), Vol. 1912. CEUR-WS.org. http://ceur-ws.org/Vol-1912/paper11.pdf

[15] Patrick J Hayes. 2009. BLOGIC or Now What's in a Link. In *Keynote presented at the 8th International Semantic Web Conference, October, Washington, DC http://videolectures. net/iswc09_hayes_blogic*. http://videolectures.net/iswc09_hayes_blogic/

[16] Urs Hunkeler, Hong Linh Truong, and Andy J. Stanford-Clark. 2008. MQTT-S - A publish/subscribe protocol for Wireless Sensor Networks. In *Proceedings of the Third International Conference on COMmunication System softWAre and MiddlewaRE (COMSWARE 2008), January 5-10, 2008, Bangalore, India*, Sunghyun Choi, Jim Kurose, and Krithi Ramamritham (Eds.). IEEE, 791–798. https://doi.org/10.1109/COMSWA.2008.4554519

[17] Luis Daniel Ibáñez, Hala Skaf-Molli, Pascal Molli, and Olivier Corby. 2013. Live linked data: synchronising semantic stores with commutative replicated data types. *IJMSO* 8, 2 (2013), 119–133. https://www.researchgate.net/profile/Pascal_Molli2/publication/264419928_Live_Linked_Data_Synchronizing_Semantic_Stores_with_Commutative_Replicated_Data_Types/links/53e8d65d0cf25d674ea87057/Live-Linked-Data-Synchronizing-Semantic-Stores-with-Commutative-Replicated-Data-Types.pdf

[18] Luis Daniel Ibáñez, Hala Skaf-Molli, Pascal Molli, and Olivier Corby. 2014. ColGraph: Towards Writable and Scalable Linked Open Data, See [22], 325–340. https://doi.org/10.1007/978-3-319-11964-9_21

[19] Leslie Lamport. 1978. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM* 21, 7 (1978), 558–565.

[20] Markus Lanthaler and Christian Guetl. 2013. Hydra: A Vocabulary for Hypermedia-Driven Web APIs. In *Proceedings of the WWW2013 Workshop on Linked Data on the Web, Rio de Janeiro, Brazil, 14 May, 2013 (CEUR Workshop Proceedings)*, Christian Bizer, Tom Heath, Tim Berners-Lee, Michael Hausenblas, and Sören Auer (Eds.), Vol. 996. CEUR-WS.org. http://ceur-ws.org/Vol-996/papers/ldow2013-paper-03.pdf

[21] Alejandro Mallea, Marcelo Arenas, Aidan Hogan, and Axel Polleres. 2011. On Blank Nodes. In *The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I (Lecture Notes in Computer Science)*, Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Fridman Noy, and Eva Blomqvist (Eds.), Vol. 7031. Springer, 421–437. https://doi.org/10.1007/978-3-642-25073-6_27

[22] Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandecic, Paul T. Groth, Natasha F. Noy, Krzysztof Janowicz, and Carole A. Goble (Eds.). 2014. *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*. Lecture Notes in Computer Science, Vol. 8796. Springer. https://doi.org/10.1007/978-3-319-11964-9

[23] Brice Nédelec, Pascal Molli, Achour Mostefaoui, and Emmanuel Desmontils. 2013. LSEQ: an adaptive structure for sequences in distributed collaborative editing. In *Proceedings of the 2013 ACM symposium on Document engineering*. ACM, 37–46. https://hal.archives-ouvertes.fr/hal-00921633

[24] Chimezie Ogbuji. 2013. *SPARQL 1.1 Graph Store HTTP Protocol*. W3C Recommendation. World Wide Web Consortium (W3C). http://www.w3.org/TR/2013/REC-sparql11-http-rdf-update-20130321/

[25] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. 2011. *A comprehensive study of convergent and commutative replicated data types*. Ph.D. Dissertation. Inria–Centre Paris-Rocquencourt; INRIA. https://hal.inria.fr/inria-00555588/document

[26] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. 2011. *Conflict-free Replicated Data Types*. Ph.D. Dissertation. INRIA. https://hal.inria.fr/inria-00609399/document

[27] Marc Shapiro, Nuno M. Preguiça, Carlos Baquero, and Marek Zawirski. 2011. Conflict-Free Replicated Data Types. https://hal.inria.fr/inria-00609399v1/document. In *Stabilization, Safety, and Security of Distributed Systems - 13th International Symposium, SSS 2011, Grenoble, France, October 10-12, 2011. Proceedings (Lecture Notes in Computer Science)*, Xavier Défago, Franck Petit, and Vincent Villain (Eds.), Vol. 6976. Springer, 386–400. https://doi.org/10.1007/978-3-642-24550-3_29

[28] Richard T. Snodgrass and Ilsoo Ahn. 1985. A Taxonomy of Time in Databases. In *Proceedings of the 1985 ACM SIGMOD International Conference on Management of Data, Austin, Texas, USA, May 28-31, 1985.*, Shamkant B. Navathe (Ed.). ACM Press, 236–246. https://doi.org/10.1145/318898.318921

[29] Ruben Verborgh, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik Van de Walle. 2014. Querying Datasets on the Web with High Availability, See [22], 180–196. https://doi.org/10.1007/978-3-319-11964-9_12

[30] Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert. 2016. Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. *J. Web Semant.* 37-38 (2016), 184–206. https://doi.org/10.1016/j.websem.2016.03.003

[31] Stephane Weiss, Pascal Urso, and Pascal Molli. 2010. Logoot-undo: Distributed collaborative editing system on p2p networks. *IEEE Transactions on Parallel and Distributed Systems* 21, 8 (2010), 1162–1174.

[32] Hafed Zarzour and Mokhtar Sellami. 2012. B-Set: A synchronization method for distributed semantic stores. In *2012 IEEE International Conference on Complex Systems (ICCS)*. IEEE, 1–6. https://www.researchgate.net/profile/Hafed_Zarzour/publication/261275016_B-Set_A_synchronization_method_for_distributed_semantic_stores/links/5559176608ae6943a876afb0/B-Set-A-synchronization-method-for-distributed-semantic-stores.pdf

Conflict Detection, Avoidance, and Resolution in a Non-Linear RDF Version Control System

The Quit Editor Interface Concurrency Control

Natanael Arndt
Agile Knowledge Engineering and Semantic Web (AKSW)
Institut für Informatik, Leipzig University
Institut für Angewandte Informatik e.V.
Leipzig, Germany
arndt@informatik.uni-leipzig.de

Norman Radtke
Agile Knowledge Engineering and Semantic Web (AKSW)
Institut für Angewandte Informatik e.V.
Leipzig, Germany
radtke@informatik.uni-leipzig.de

ABSTRACT

The Semantic Web is about collaboration and exchange of information. While the data on the Semantic Web is constantly evolving and meant to be collaboratively edited there is no practical transactional concept or method to control concurrent writes to a dataset and avoid conflicts. Thus, we follow the question, how can we ensure a controlled state of a SPARQL Store when performing non transactional write operations? Based on the Distributed Version Control System for RDF data implemented in the Quit Store we present the Quit Editor Interface Concurrency Control (QEICC). QEICC provides a protocol on top of the SPARQL 1.1 standard to identify, avoid, and resolve conflicts. The strategies *reject*, *branch*, and *merge* are presented to allow different levels of control over the conflict resolution. While the *reject* strategy gives full control to the client, with *branch* and *merge* it is even possible to postpone the conflict resolution and integrate it into the date engineering process.

CCS CONCEPTS

• **Information systems → Query languages**; **Version management**; **Resource Description Framework (RDF)**; *Data management systems*; *Middleware for databases*; RESTful web services.

KEYWORDS

versioning, transaction, concurrency control, conflict detection, Git, RDF, Quit Store

ACM Reference Format:
Natanael Arndt and Norman Radtke. 2019. Conflict Detection, Avoidance, and Resolution in a Non-Linear RDF Version Control System: The Quit Editor Interface Concurrency Control. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316519

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316519

1 INTRODUCTION

Collaboration and exchange of information is deep-seated in our culture. With the Web, mankind has built a global communication system based on the exchange of data. Especially, the publication and exploitation of Linked Data is gaining popularity as it is apparent in the *Linked Open Data Cloud* with 1,234 datasets and 16,136 links between the datasets[1] also known as the Semantic Web. On the Semantic Web knowledge bases and ontologies are used to encode a common understanding of people on a distributed network. Collaborative editing of a common knowledge base is an evolutionary process of creating new versions of the knowledge base. In order to manage collaboration on Linked Data we face the problem of distributed versioning and parallel access to the knowledge base. Recently, the versioning of RDF data has gained attention [6, 7]. Evolution of data and contributions by distributed parties introduce the problem of overlapping update operations, called concurrency. The problem of concurrency control was extensively studied in the context of conventional databases and collaborative file exchange systems [4, 5, 8, 20, 21]. But the architecture of Semantic Web Applications is mostly Web oriented and similar to the architecture of Ajax/REST Web Applications. Web Applications as well as Semantic Web Applications consist of three layers: *persistence layer*, *data interchange & transaction processing*, and *user interface* [16]. The problem of distributed consistency is studied with regard to the transfer of distributed states in Web Applications [11, 15, 19].

However, the role of the persistence layer and thus the data interchange in Semantic Web Applications is different to the architecture of conventional applications. On the informal *Semantic Web Layer Cake*[2,3] the bottom layers cover the abstract representation to encode data as triples and their serialization (cf. *persistence layer* [16]). To query the data model the SPARQL 1.1 Query and Update Language is standardized [10, 12] (cf. *data interchange & transaction processing*). The top most layer represents *User Interface & Applications* that are used to read and change the data. The RDF data model allows Semantic Web Applications to encode the schema and the application logic within the abstract representation on the persistence layer [13, 14, 16]. This semantically rich data, stored in an abstract data model allows to handle concurrency on an abstract level while providing more flexibility to resolve conflicts.

[1] http://lod-cloud.net/; As of June 2018
[2] https://www.w3.org/2007/03/layerCake.svg
[3] https://natanael.arndt.xyz/notes/semantic-web-layer-cake

But, only little research is performed with regard to the management of concurrent updates on RDF data. For SPARQL 1.1 (cf. [22]) there is no standardized transaction interface available even though several approaches[4,5] exist. The related work mainly tries to adapt the protocols from conventional databases [17, 18] to SPARQL 1.1 by striving for serializability. Innovative concepts that take the possibilities of distributed versioning systems and the abstract concept of RDF knowledge bases into account are missing.

In this paper we present the Quit Editor Interface Concurrency Control (QEICC), a method to handle conflicting write operations on a SPARQL 1.1 Update interface of a distributed multiversion RDF quad store. The QEICC is build on top of *Quit*, our previously published distributed version control system for RDF knowledge bases [2, 3]. The Quit Store maintains multiple *diverged* version logs (*branches*) of an RDF dataset while each branch can represent a different state and latest version of the dataset. The branched versioning logs allow to postpone the *reconciliation* of diverged versions and to incorporate it into the data engineering process. Further, it is possible to retrieve any version on the Quit Store using its random access feature. With the QEICC extension we exploit the properties of the distributed version control system of the Quit Store. Using the branching system we postpone the conflict resolution as we are not constrained to reach serializability. The reconciliation system is used to still strive for an eventually conflated and serialized versioning log. In addition we exploit the commit system and the random access to arbitrary versions of the Quit Store to provide snapshot isolation for the operations. As we use these features, with QEICC we detect, avoid, and resolve conflicts of update operations based on a snapshot identification. To give full control to editing interfaces within the data engineering process we provide the three optimistic concurrency control strategies *reject*, *branch*, and *merge*. The strategies differ in their conflict handling which allows us to shift the resolution of a conflict between the data's creator, automatic processes, and collaborators to perform a data-centric reconciliation.

The remainder of this paper is structured as follows. First, we give an overview on the state of the art for data base systems and Web Applications as well as Semantic Web Applications in section 2. Second, we provide a problem description in section 3. The main contribution of our work is the Quit Editor Interface which is presented in section 4 with the three conflict resolution strategies *reject* in section 4.1, *branch* in section 4.2, and *merge* in section 4.3. The strategies are compared regarding their individual properties and the instance of control in section 5. Finally, we conclude the paper in section 6.

2 STATE OF THE ART

In the following we take a look at the related work. First we will provide a historical walk-through on conflict detection, concurrency, and transaction management for data base systems and derived concepts for RDF data stores in section 2.1. Afterwards we look into approaches for state transfer between clients and data stores respective back-ends in the application engineering area and in particular for Semantic Web Applications in section 2.2.

[4]https://jena.apache.org/documentation/txn/
[5]http://people.apache.org/~sallen/sparql11-transaction/

2.1 Database Management Systems

The problems of *concurrency control* and the *recovery problem* were broadly discussed by Bernstein et al. [5].

> Systems that solve the concurrency control and recovery problems allow their users to assume that each of their programs executes atomically – as if no other programs were executed concurrently – and reliably – as if there were no failures.

This abstraction is called *transaction* and an algorithm that executes the transactions atomically is called *concurrency control algorithm*. The concurrency control algorithm is implemented by executing concurrent interleaving transactions one after another *to give the illusion that transactions execute serially*. Interleaving executions of transactions that have the same effect as serial executions are called *serializable*, which is considered correct because they support the *illusion of transaction atomicity*.

Berenson et al. [4] introduce the *snapshot isolation* type. This isolation type allows reads or writes to be executed on a snapshot which ensures isolation from other transactions but is not serializable. Following from the isolation guarantee snapshot isolated write operations can only be performed on distinct data items.

The problem to ensure consistency across all replicas while clients are disconnected from the network was discussed for the *Coda* file system by Satyanarayanan et al. [21]. If a client is disconnected, be it intentionally or not, the client can still perform local changes in the cached file system. When the client is reconnected a reintegration process starts that tries to execute the cached updates on the file system of all replicas. If conflicts between updates are detected the conflicting files are temporarily stored in a *covolume* until the conflict is resolved by a user.

Demers et al. [8] present the *Bayou System*, a client/server platform to replicate databases for the usage on mobile devices like *personal digital assistants*. The system propagates all writes through a peer-to-peer protocol to a *primary* server which accepts writes to be *committed*. All other write operations on *secondary* servers that were not yet committed are *tentative* until they are serialized on the *primary* server and the order of writes propagated to the secondary servers. If conflicts arise *mergeprocs* are employed that understand the semantics of the data format and domain and can reconcile the conflict.

Another system to reconcile transactions performed on mobile clients is presented by Phatak and Badrinath [20]. Similar to the *Bayou System* all transactions need to be transmitted to a server to be globally committed. If a disconnected client performs local transactions, on reconnection they need to be tested for serializability and are rolled back on conflicts. Further they discuss the weakening of the serializability guarantee which can be enabled by a semantic aware reconciliation and by providing snapshot isolation (cf. [4]). To implement the multiversion reconciliation the conflict detection and resolution is decoupled, while the server is responsible for the conflict detection the client provides the conflict resolution function.

Muys [17] provides a discussion about a concurrency control protocol for multiversion RDF data stores. In contrast to relational or object databases in an RDF store the smallest "cell" that can be considered is the entire graph. If multiple graphs are present

and can be targeted by an update, the whole RDF dataset needs to be considered as a unit of change. This results in a single global write lock when applying traditional concurrency control protocols on an RDF data store and thus provides very little concurrency. To avoid a single global write lock the usage of predicate locking is proposed with the basic graph pattern (BGP) as predicate that is tested against the update set of a transaction that needs to be serialized. Still the resolution of conflicts is a problem. For this purpose the *single version optimistic concurrency control* protocol is adapted to a multiversion store by specifying the *multiversion optimistic concurrency control* protocol. The multiversion protocol allows to exploit snapshot isolation and thus to execute read-only operations from the validation protocol.

Neumann and Weikum [18] are following a similar approach and extend the RDF-3X store to x-RDF-3X. The RDF store provides versioning with the ability for time travelling queries. The store supports snapshot isolation and full serializability with help of predicate locks based on the query BGPs. Due to the snapshot isolation no locks are needed for read operations. To avoid too small locks or unnecessary large locks a lock splitting algorithm is employed. Update operations are performed in a per-transaction workspace that is merged into the differential indexes if a savepoint is issued.

2.2 Application Engineering

Ousterhout and Stratmann [19] discuss the state problem in Web Applications using asynchronous AJAX requests. They analyze and verify possible solutions for managing states on the client side or on the server side in a Web Application. The authors conclude that both, client side and server side states have drawbacks. Client side state management has *overheads for shipping state between browser and server, and it creates potential security loopholes by allowing sensitive server state to be stored in the browser.* Whereas the server-based approach *introduces overheads for saving state as part of sessions, and it has garbage-collection issues that can result in the loss of state.* The authors prefer the usage of a server-based state management over that of a browser-based and predict upcoming challenges in state management of Web Applications.

Pardon and Pautasso [11] specify a protocol to support atomicity and recovery over distributed REST resources. The authors contribute to the debate in the REST community to whether or not transaction support is needed. The presented approach references and is similar to the position paper by Helland [15]. The need for the protocol is motivated by a business use case on booking two connecting flights from two different airlines. The presented approach is based on a Try-Cancel/Confirm (TCC) pattern using (a) an initial state, (b) a reserved state and (c) a final state. The reserved state (b) is called *tentative* in [15]. To form a transaction an arbitrary number of REST services are loosely coupled. A transaction is valid if all reserved states (b) are confirmed, after this the final state (c) is entered for each service. Compared to the two-phase commit lock the authors point out, that the TCC-approach *offers higher-level semantics and does not hold low-level database locks.* The participants *do not block any other work other than the one affected by the business resources they reserve.* Both approaches [11, 15] mention that an uncertainty is left if one of the loosely coupled systems fails before all systems have reached the final state. The management of uncertainty must be implemented in business logic to counter this issue.

Web Applications as well as Semantic Web Applications (SWA) consist of three layers *Persistence Layer*, *Data Interchange & Transaction processing*, and *User Interface* [16]. Martin and Auer [16] present a categorization model for Semantic Web Applications. Following this categorization model our focus is on intrinsic and extrinsic producing SWAs while the level of user involvement, semantic richness, and semantic integration is left open. In an empirical study performed by Heitmann et al. [13, 14] all SWAs have a graph access layer, an increasing amount of SWAs is concerned with data creation, and some applications even provide structured data authoring interfaces. The data flow to allow collaborative knowledge acquisition, visualization, and creation requires for a management of concurrent and possibly conflicting operations. The above presented research on non-semantic Web Applications is mostly concerned with distributed state management. Methods like Flux[6] can be used in client applications to bundle the data flow towards the back-end within an application. A store object ensures that all view component's requests to the back-end system are executed and controlled in a single place which can represent a client side state. While the issue of communicating the global state of the store between back-end and client application is underrepresented in research. A reason for a lack of research in this are in the Ajax/REST community might be the fact, that Ajax/REST applications usually encode parts of the business logic in the back-end. In contrast to that, SWAs strive to separate the data's semantics and implementation by searching for a generic semantic representation. With the RDF data model SWAs are able to encode semantics in the data store and thus are able to directly communicate with the RDF store as persistence layer.

3 PROBLEM DESCRIPTION

In user facing applications, in particular in Web Applications, the user interface is constructed of many components. Each of the user interface components might send read requests to the back-end system. When the user performs any action that induces an update operation the user might take into account any information displayed in the user interface for his update decision. These requests are sent by the individual components or can be coordinated by a method like Flux. In many traditional systems the back-end system implements parts of the business logic and can thus manage the state of the front-end in interaction with the store object. For a Semantic Web Application the data store can directly serve as back-end and be accessed using the SPARQL 1.1 Protocol [9]. In the following we demonstrate two scenarios of clients interacting with a SPARQL 1.1 store to manage a todo-list.

```
@prefix ex: <http://example.org/> .
ex:garbage a ex:Todo;
  ex:task "Take out the organic waste" .
```

Listing 1: The initial data in the triple store to be considered for scenarios 1 and 2.

[6] https://facebook.github.io/flux/docs/in-depth-overview.html#structure-and-data-flow

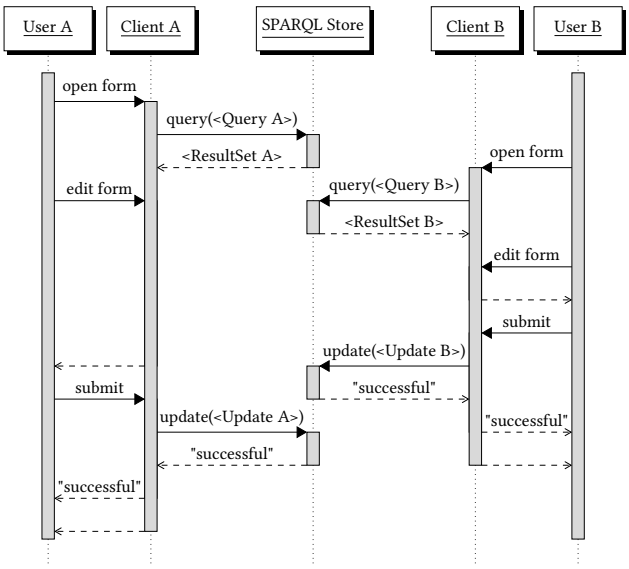


Figure 1: Sequence diagram of two overlapping update operations which result in an uncontrolled state of the shared data store.

```
PREFIX ex: <http://example.org/>
INSERT {
  ?task ex:status ex:completed
}
WHERE {
  ?task a ex:Todo
}
```

Listing 2: SPARQL Update operation *<Update A>* to mark the task as completed in scenario 1.

```
PREFIX ex: <http://example.org/>
INSERT DATA {
  ex:chain a ex:Todo;
    ex:task "Lubricate␣the␣bike␣chain."
}
```

Listing 3: SPARQL Update operation *<Update B>* to add a new task in scenario 1.

Scenario 1. In fig. 1 an example of two conflicting operations based on updating data through a form is depicted. We assume the content of the SPARQL Store are the statements shown in listing 1. Both clients retrieve the data from the store with a SPARQL Query and receive the respective result set containing the task to do. While user A performs the single task in the store and changes it to "completed" with the update operation in listing 2, meanwhile user B adds a new task with the update operation in listing 3. In the final state of the store both tasks will be marked as completed. If the operations were executed in the opposite order the state of the store would be different. Thus the two operations are not serializable and would cause the operations to abort with failure in traditional transactional concepts. But the operations do not fail in current SPARQL 1.1 implementations.

Scenario 2. Let us consider another scenario following the execution of fig. 1 and with the assumed initial content of the SPARQL Store as shown in listing 1. Both clients retrieve the data from the store with a SPARQL Query and receive the respective result set containing the task to do. While user A performs the task `ex:garbage` and changes it to "completed" with the update operation in listing 4, user B has changed the task description with the update operation in listing 5. Even though both update operations have taken care to identify the data item to change, in the final state of the store the changed task will be marked as completed even though this is not correct. Further, also the execution of the operations in the opposite order results in the same result. A traditional transactional concept could not identify the operations as conflicting.

```
PREFIX ex: <http://example.org/>
INSERT DATA {
  ex:garbage ex:status ex:completed
}
```

Listing 4: SPARQL Update operation *<Update A>* to mark the task as completed in scenario 2.

```
PREFIX ex: <http://example.org/>
DELETE { ?todo ex:task ?task }
INSERT {
  ?todo ex:task "Take␣out␣the␣organic␣waste␣
    ↪ and␣the␣residual␣waste"
}
WHERE {
  BIND (
    "Take␣out␣the␣organic␣waste" as ?task)
  ?todo ex:task ?task
}
```

Listing 5: SPARQL Update operation *<Update B>* to change the description of the task in scenario 2.

For SPARQL 1.1 no transactional concept is standardized. The SPARQL 1.1 Update language allows to build update operations of the form `DELETE {} INSERT {} WHERE {}` [10]. The `WHERE` part can be used to encode preconditions for the update operations and bind variables used in the update operations, as shown in the listing 2 and 5. But enclosing all preconditions inferred from all read operations that were used to compose the user interface is not a practical way to ensure it is not conflicting with other operations. Further more, write operations in SPARQL 1.1 can also be performed without any precondition (`INSERT DATA`, `DELETE DATA`). Our question is: how can we ensure a controlled state of the SPARQL Store when performing non transactional write operations?

4 THE QUIT EDITOR INTERFACE

In the following we present the Quit Editor Interface Concurrency Control (QEICC) a method to handle concurrency of write operations in a distributed multiversion RDF quad store. QEICC is built

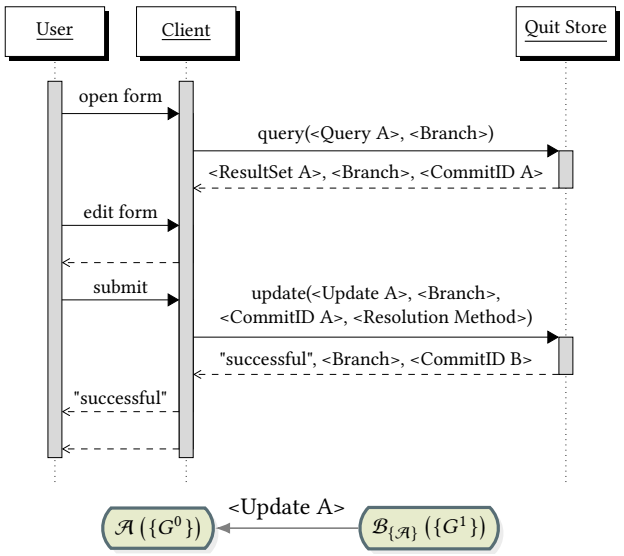


Figure 2: Sequence diagram showing the execution of an update operation with the parameters specified by the Quit Editor Interface. In the lower part, the commit graph is depicted which exists at the Quit Store after the update operation.

on top of the Quit Store [2, 3]. We first introduce the essential preliminaries of the Quit concept. Then, we propose a fully backward compatible[7] protocol on top of the SPARQL 1.1 Protocol [9] that allows us to control concurrency. Based on this protocol we present the three conflict resolution strategies *reject* in section 4.1, *branch* in section 4.2, and *merge* in section 4.3.

The QEICC builds on our previously published distributed version control system (DVCS) for RDF knowledge bases Quit [2, 3]. The underlying DVCS allows not only linear versioning logs but also non linear *diverged* versioning logs represented by an acyclic directed graph of multiple *branches*. Because of the non linear versioning log we are not constrained to reach serializability of all performed operations. Instead, the store maintains multiple branched versioning logs while each branch can represent a different state and latest version of the database. Each version in the versioning log is identified by a unique hashed *commit id*, further each version references its predecessor as parent commit. A *branch* is identified by its name and is a pointer to the currently latest commit id in its versioning log. The store allows random access to arbitrary versions in the versioning log for query and update operations. Thus it is possible to start a new diverged versioning log at any existing version in the log. The latest version (*Head*) of each version log respective branch can be queried and updated through virtual SPARQL Endpoints. The store provides one virtual SPARQL Endpoint for each branch, i.e. http://localhost:5000/sparql/<Branch>.

The QEICC method exploits the commit system and the random access feature of the underlying DVCS to provide a state identification of the data store to the client. As shown in fig. 2 the state is transferred on query and update operations. A query operation is sent to the store following the standard SPARQL 1.1 Protocol [9]. The client selects the branch to work on by using the respective SPARQL Endpoint for the branch. In fig. 2 the selection of the branch is depicted by the parameter *<Branch>* of the *query()* and *update()* methods.

When the client sends a query operation to the store, the result set for the query is returned along with the additional two HTTP-Headers `X-CurrentBranch` and `X-CurrentCommit` (*<Branch>* and *<CommitID A>* in fig. 2). The `X-CurrentBranch` header field contains the name of the branch that was requested. The `X-CurrentCommit` header field contains the commit id of the current Head of the selected branch in the store. With these two header fields the store encodes the current snapshot state of the store at the time of the execution of the query.

When the client sends an update operation to the store it transmits the snapshot state that it has received with the last query (`X-CurrentCommit`). To transfer the state from the client to the store the client sets the parameter `parent_commit_id`[8] along with the update operation to the store (*<CommitID A>* in fig. 2). This transfer of the state allows the client to express towards the store which is the currently most recent commit on the branch. Additionally the client selects a resolution strategy that the store has to apply in case of a conflict. The client sets the resolution method with the `resolution_method` parameter along with the `parent_commit_id` and update operation (*<Resolution Method>* in fig. 2). The `resolution_method` is one of the values `reject`, `branch`, or `merge`. When the store receives an update operation it compares the commit id of the currently latest commit (*Head*) on the branch with the value of `parent_commit_id` which was sent by the client along with the update request. By comparing the commit ids the store can identify whether a conflict exists or not.

No conflict. If no overlapping update operation was performed in the meantime both commit ids are equal. In this case the store commits the update operation and returns the commit id of the new Head of the selected branch. The now updated commit graph is depicted in the lower part of fig. 2. The state before the update is encoded in commit $\mathcal{A}$ with graph G^0, the update is performed using the update operation *<Update A>* which leads to the new state $\mathcal{B}_{\{\mathcal{A}\}}$ which was derived from $\mathcal{A}$ and points to its predecessor (for more details cf. [1, 2]).

Conflict. If the commit ids are different a conflict is detected. This case is depicted in fig. 3. The conflict is then handled according to the specified value for the `resolution_method`. To give control over the conflict resolution process we provide three optimistic strategies which differ in their conflict handling. In fig. 3 the previous queries and the conflicting update operation are enclosed in the *Overlapping Edit* block. For brevity, this block is not repeated but referenced in the later figures, fig. 4 to 6.

4.1 Reject

If a conflict is detected on the server side, choosing the *reject* strategy will abort the execution of the operation and will leave the

[7] The protocol still allows standard SPARQL 1.1 Query and Update operations, which of course will not benefit from the concurrency control features.

[8] The client uses HTTP form-encoded or query string request parameters to encode the values.

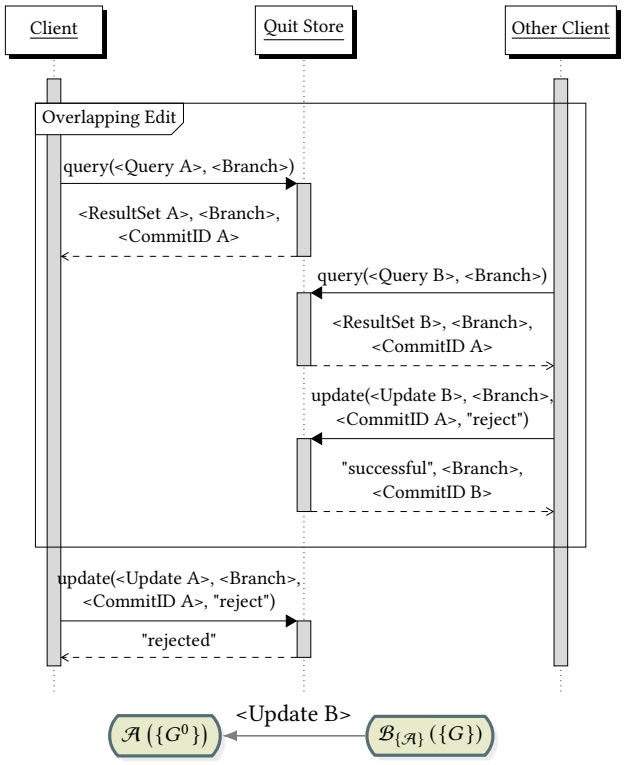


Figure 3: Sequence diagram showing two overlapping update operations, where the first operation is committed, while the second operation is aborted according to the selected strategy *reject*. The commit graph shows the state of the store after the successful update operation was committed.

store unchanged. This allows to perform update operations in a try-catch or trial-and-error manner. Figure 3 shows a rejected update operation performed by the client. The commit log in the lower part of fig. 3 shows the state of the store after the successful commit of the update operation from the other client. After the rejected operation it is up to the clients implementation how to proceed.

One possibility to resolve the conflict on the client side is depicted in fig. 4. The *Overlapping Edit* block is reused by reference from fig. 3. To resolve the conflict this method involves a new retrieval of the data that is been edited on the client side. Just after the update operation was rejected the query operation is performed again and returns now an up-to-date result set *ResultSet A'*. This result set is compared to the original result set in two alternative blocks. The top block is for the case that the result set of the query operation did not change and the bottom part for if the result set changed. If the result set is the same as the original result set the update operation is resent now with the updated commit id as parameter. By comparing the result sets we can ensure that the operation with which we are conflicting did not change the part of the graph (subgraph) that was queried with our query operation. This involves the assumption that the update operation was constructed based on the result set retrieved by the query operation *Query A* (cf. fig. 3). If the update

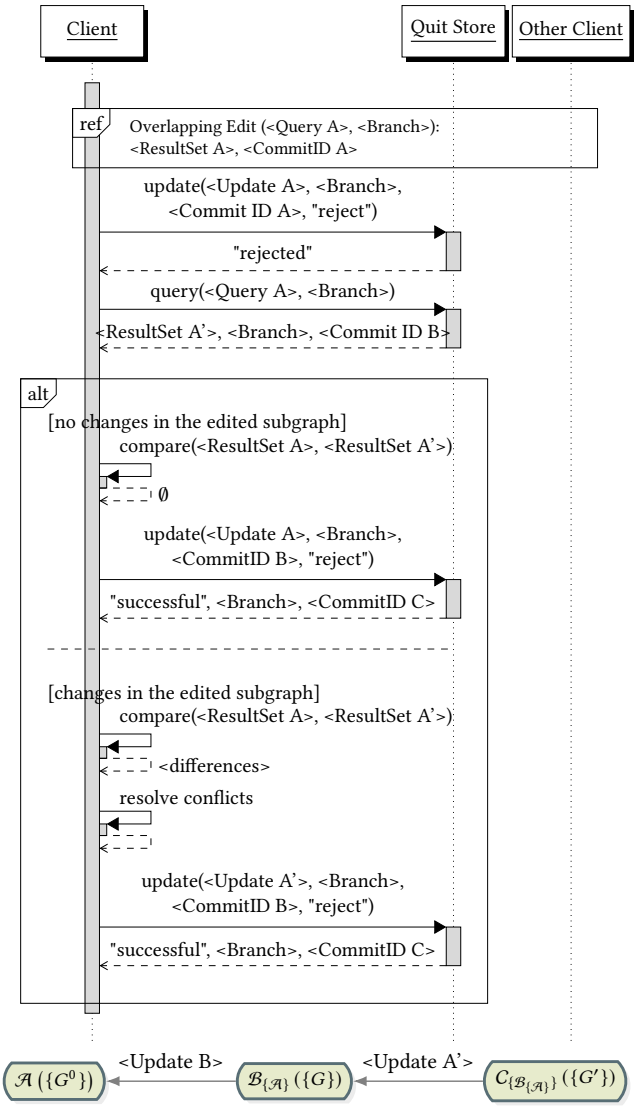


Figure 4: Sequence diagram showing the retry of an update operation after it was rejected due to an overlapping update operation. The alternative flows depict the first case that the result set of the original query did not change and thus the update operation can be resent. The second case in which the client needs to perform a resolution of the conflict before it can resent an updated operation. The commit graph shows the state of the store after the client has performed a resolution and the new update operation was committed.

affected the subgraph of our interest the client needs to resolve the conflict locally. This resolution can happen by informing the user about the changes and let the user review its input under the new circumstances and resubmit the form. After the conflict was resolved on the client side the adapted update operation can be submitted to the store along with the new commit id. The bottom part of fig. 4 shows the version log with the update operation by the

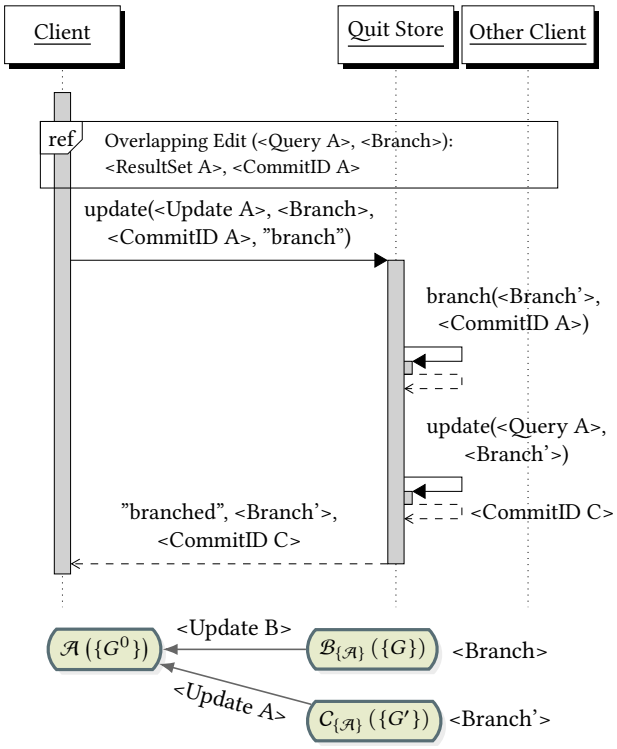


Figure 5: Sequence diagram of the overlapping update operation with the resolution strategy *branch*. The update operation is committed to a new branch that can be used to postpone the resolution. The commit graph shows how the state of the store has diverged into two branches after committing both update operations.

other client (as shown in fig. 3) and the adapted update operation after the client side conflict resolution and the new commit C.

4.2 Branch

The *branch* strategy actually exploits the possibilities of our non-linear version log system. Figure 5 shows the execution of an update operation that conflicts and is resolved with the branch strategy. In contrast to the other two strategies the branch strategy does never fail. Instead, if a conflict arises a new branch is created that diverges from the original branch from the last common commit before the conflict on the branch. This divergence is shown in the lower part of fig. 5, each of the commits $\mathcal{B}$ and C are derived from the same original commit $\mathcal{A}$. In the sequence diagram the new branch is created with the *branch()* method. After the new branch was created the conflicting update operation is committed to the new branch *<Branch'>*.

This strategy allows fire-and-forget operations in a way that the update operation is committed, no matter if it conflicts or not. Also, this strategy allows to postpone a conflict resolution for instance if the correct resolution is not yet known or needs to be found based on a community discussion. This allows a data-centric reconciliation, in contrast to time based reconciliation, of the conflict using appropriate, e.g. domain specific, merge operations (cf. section 4.3 and [2]). Another asset of allowing a versioning log to branch is that it reflects the actual lineage of the update operations and changed dataset. By analysing the version log it is possible to reconstruct in which logical order operations were performed.

4.3 Merge

The *merge* strategy is similar to the branch strategy as it creates a new branch to commit the update operation when a conflict arises. In addition it subsequently performs a reconciliation of the diverged versioning log. To perform the merge, a merge method can be selected according to the methods supported by the respective system. The Quit Store has specified the merge methods *Three-Way-Merge* and *Context Merge* [2]. The two possibilities for a successful and a failing merge process are depicted in the two alternative blocks. If the reconciliation is successful a new commit is created as shown in the lower part of fig. 6, the commit $\mathcal{D}$ is derived from the two diverged commits $\mathcal{B}$ and C and combines the two branches again. If the reconciliation fails the update operation fails and the system ends up in the branched state. This means that the update operation is committed but the diverged states could not be automatically reconciled. The result of a failed merge operation is the description of the *merge conflicts* and the references to the new branch and commit. The merge conflict now needs to be resolved by the client or can be postponed to a later state. If the merge conflicts are resolved a new *merge* operation is performed by the client.

A merge method compares the two versions of the dataset as they are produced by the conflicting operations. This comparison usually takes into account the last common ancestor of the two newly created commits. In our case the newly created commits are $\mathcal{B}$ and C and the last common ancestor is $\mathcal{A}$. Each of the datasets $\{G\}$ and $\{G'\}$ is compared to the dataset $\{G^0\}$, this allows to break down the conflict from the level of overlapping operations to the actual data in the graphs. Besides the predefined merge methods *Three-Way-Merge* and *Context Merge* additional methods can be implemented to reflect special properties of a domain model. This allows a semi-automatic data-centric reconciliation of overlapping operations.

5 COMPARISON

The presented strategies have different properties with regard to the control over the resolution process. The resolution strategies are operating on a DVCS which provides non-linear versioning logs. Due to the distribution and because we are not limited to work on a linear versioning log we gain flexibility. The flexibility allows us to distribute the responsibility to resolve a conflict and to freely define a point in time at which a conflict resolution has to be performed.

In table 1 the three strategies reject, branch, and merge are categorized according to the aspects *performer of the resolution*, *point in time*, and *on failed resolution*. In the case that an operation is detected to conflict with a previous operation the server performs the strategy that was specified by the client through the parameter `resolution_method`. For the *reject* strategy the server rejects the operation and thus gives the control back to the client or user to perform the resolution. The client can consequently decide at which point in time the conflict should be resolved, if also the resolution

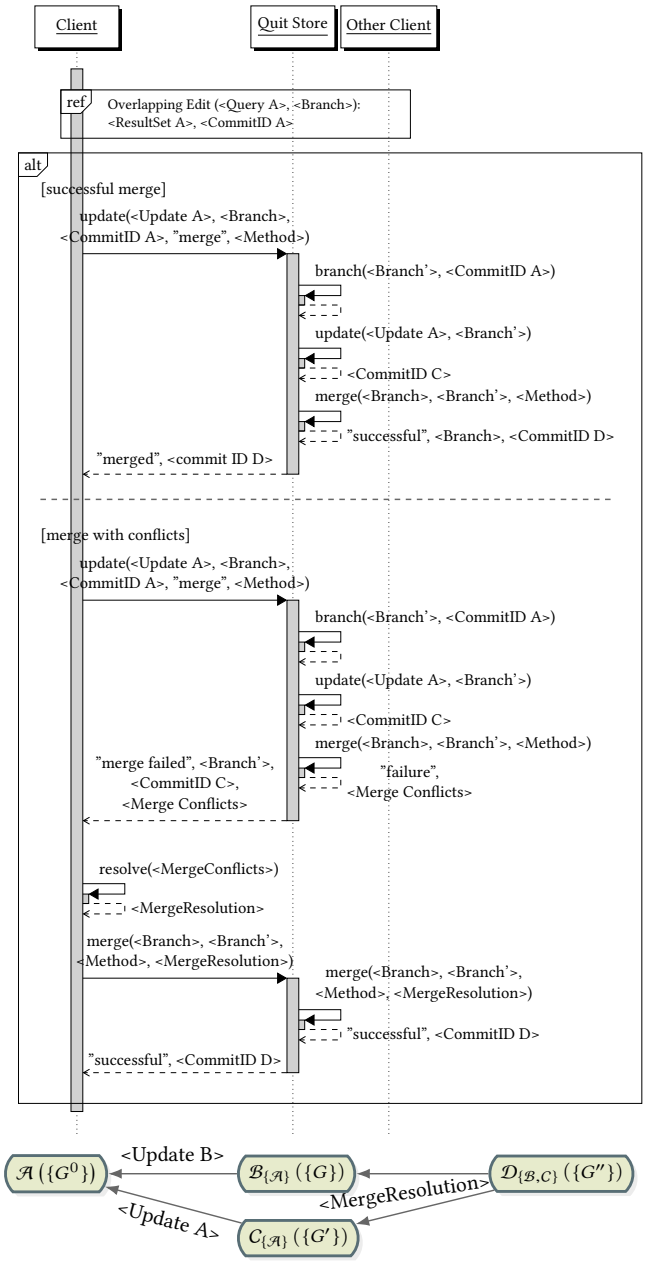


Figure 6: Sequence diagram of the overlapping update operations and server side resolution by merging the changes. The merge is successful in the first case, while in the second case the merge conflicts are reported to the client and need to be resolved on the client side. The commit graph shows the merged state of the store after a successful merge operation was performed with an optional *<MergeResolution>*.

fails the client can retry or abort. With the method *branch* a new branch is created that is part of the non-linear versioning log, which enables any collaborator in the DVCS to perform a conflict resolution on this branch. The resolution of a branched conflict can be

Table 1: Comparison of the three conflict resolution strategies *reject*, *branch*, and *merge*.

Strategy	Performer of the Resolution	Point in Time	On Failed Resolution
reject	client/user	client's choice	retry
branch	any collaborator	postponed	n/a
merge	merge method	immediate	resolve/branch

postponed to any later point in time. The *merge* strategy relies on a merge method that performs a predefined resolution procedure and can thus automatically control the resolution of the conflict. The merge method is applied immediately after the conflicting update operation was applied on a temporary branch. A successful conflict resolution is not always guaranteed by the merge method, depending on its implementation. In case of a failed resolution the merge method gives the control back to the client in the form of a *merge conflict*. If the client can resolve the merge conflict the resolution was successful, if not we end up at a branched state whose resolution is postponed in the same way as for the branch strategy.

In summary, the control of the resolution for the *reject* strategy is the client's responsibility, for the *branch* method the control can be distributed among collaborators, and for the *merge* strategy the control is at an automatic merge method but given back to the client or the collaborators in case of failure. We thus can shift the resolution of a conflict between the data's creator, an automatic process, and collaborators. The collaborator to resolve the conflict can be any participant in a team or a specified role according to a data engineering workflow. In contrast to traditional concurrency control systems when we deal with RDF data, we are able to incorporate the semantics of the data into the resolution process. In different usage scenarios different stakeholders, agents, or collaborators are responsible or capable to interpret the semantics of the data. With the presented strategies we gain flexibility to provide the control of a conflict resolution to the respective role in charge.

6 CONCLUSION

The usage of the SPARQL 1.1 Query, Update, and Protocol standards allows interoperability across software stacks in different usage scenarios. But this standard is not meant to deal with problems occurring in collaborative scenarios. The overall problem is how can we ensure a controlled state of the SPARQL Store when performing non transactional write operations. In the collaborative usage scenario we thus have additional requirements on top of the functionality covered by the SPARQL 1.1 standard.

With the Quit Editor Interface Concurrency Control (QEICC) we have presented a method to identify and avoid overlapping update operations respective conflicts. The conflicts can not only be detected on operations that would not be serializable in traditional transaction protocols, but also in cases where the operations could be executed in an isolated way. This is achieved by the possibility for the client to express towards the store which state it assumes to be the most recent respective which is the currently most recent commit on a branch. On top of this we have defined three strategies to handle the conflicts. The presented system allows to strive for

a linear execution of the operations and only diverges if overlapping operations are detected. In contrast to conventional systems our system is based on non-linear versioning logs. The branched versioning logs allow us to postpone the serialization of operations (*reconciliation*) and incorporate it into the data engineering process. Due to the different strategies provided, the control over the resolution and reconciliation process can be assigned to the respective roles. The reconciliation can happen in a data-centric way that is aligned with the domain model.

As a side-effect, using the parameters defined by QEICC for conflict detection on updates the client can store the received values of the current branch and current commit id as the state of the store. This state representation allows the client to detect updates to the store while it performs query operations. This detection allows the client to trigger local updates of the user interface components when the store changes, to always present up-to-date information.

The QEICC mainly focuses on conflicts on local instances of a Distributed Version Control Systems for RDF data i.e. the Quit Store. The Quit Store provides, besides its local branching features, the possibility to run globally distributed networks of data repositories. Between the instances there is no steady connection, but committed update operations can be distributed among the instances via a synchronization protocol. This synchronization protocol transmits the individual commits as well as the branches. This synchronization of branches allows to even reconcile update operations that happened at different locations in the same way as following the *branch* strategy.

With our Quit system we hope that Distributed Version Control Systems find their way into the RDF data engineering and knowledge engineering domain to allow more collaborative and agile processes. With QEICC on top we provide an interface to control overlapping update operations and assign the resolution task to the responsible entity. Due to the abstract conception of the QEICC method it could also be adapted to RDF archiving and linear versioning systems. The `X-CurrentCommit` header field can be set to any string that distinctively identifies a state of the graph in the version log. The `X-CurrentBranch` header field can be set to a fixed value. This would allow to implement the reject strategy but also the merge strategy can be implemented for successful merging by performing the branch and merge operations in the cache. While this would not allow to postpone the reconciliation, it provides the possibility to detect conflicts and provides the flexibility of domain specific merge methods.

ACKNOWLEDGMENTS

This work was partly supported by a grant from the German Federal Ministry of Education and Research (BMBF) for the LEDS Project under the grant number 03WKCG11C, by the Federal Ministry of Transport and Digital Infrastructure (BMVI) for the LIMBO project under the grant number 19F2029G, and the Federal Ministry for Economic Affairs and Energy (BMWi) for the Platona-M project under the grant number 01MT19005A.

REFERENCES

[1] Natanael Arndt and Michael Martin. 2017. Decentralized Evolution and Consolidation of RDF Graphs. In *17th International Conference on Web Engineering (ICWE 2017)*. Rome, Italy. DOI: http://dx.doi.org/10.1007/978-3-319-60131-1_2

[2] Natanael Arndt, Patrick Naumann, Norman Radtke, Michael Martin, and Edgard Marx. 2018. Decentralized Collaborative Knowledge Management using Git. *Journal of Web Semantics* (2018). DOI: http://dx.doi.org/10.1016/j.websem.2018.08.002

[3] Natanael Arndt and Norman Radtke. 2017. A Method for Distributed and Collaborative Curation of RDF Datasets Utilizing the Quit Stack. In *INFORMATIK 2017*. Chemnitz, Germany. DOI: http://dx.doi.org/10.18420/in2017_187

[4] Hal Berenson, Philip A. Bernstein, Jim Gray, Jim Melton, Elizabeth J. O'Neil, and Patrick E. O'Neil. 1995. A Critique of ANSI SQL Isolation Levels. In *SIGMOD Conference*.

[5] Philip A. Bernstein, Vassos Hadzilacos, and Nathan Goodman. 1987. *Concurrency Control and Recovery in Database Systems*. Addison-Wesley.

[6] Jeremy Debattista, Javier D Fernández, Jürgen Umbrich, and Maria-Esther Vidal. 2018. Preface of MEPDaW 2018: Managing the Evolution and Preservation of the Data Web. In *MEPDaW 2018*.

[7] Jeremy Debattista, Javier D Fernández, Maria-Esther Vidal, and Jürgen Umbrich. 2019. Managing the evolution and preservation of the data web. *Journal of Web Semantics* (2019). DOI: http://dx.doi.org/10.1016/j.websem.2018.12.011 editorial.

[8] Alan J. Demers, Karin Petersen, Mike Spreitzer, Douglas B. Terry, Marvin Theimer, and Brent B. Welch. 1994. The Bayou Architecture: Support for Data Sharing Among Mobile Users. *First Workshop on Mobile Computing Systems and Applications* (1994).

[9] Lee Feigenbaum, Gregory Todd Williams, Kendall Grant Clark, and Elias Torres. 2013. *SPARQL 1.1 Protocol*. Recommendation. W3C. https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/

[10] Paula Gearon, Alexandre Passant, and Axel Polleres. 2013. *SPARQL 1.1 Update*. Recommendation. W3C. https://www.w3.org/TR/2013/REC-sparql11-update-20130321/

[11] Cesare Pautasso Guy Pardon. 2011. Towards Distributed AtomicTransactions over RESTful Services. (2011). DOI: http://dx.doi.org/10.1007/978-1-4419-8303-9_23

[12] Steve Harris and Andy Seaborne. 2013. *SPARQL 1.1 Query Language*. Recommendation. W3C. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/

[13] Benjamin Heitmann, Richard Cyganiak, Conor Hayes, and Stefan Decker. 2012. An Empirically Grounded Conceptual Architecture for Applications on the Web of Data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* (2012).

[14] Benjamin Heitmann, Richard Cyganiak, Conor Hayes, and Stefan Decker. 2014. Architecture of Linked Data Applications. In *Linked Data Management*.

[15] Pat Helland. 2007. Life beyond Distributed Transactions: an Apostate's Opinion.. In *CIDR*.

[16] Michael Martin and Sören Auer. 2010. Categorisation of Semantic Web Applications. In *proceedings of the 4th International Conference on Advances in Semantic Processing (SEMAPRO2010)*. Florence, Italy. http://svn.aksw.org/papers/2010/SEMAPRO_Categorisation_SWA/public.pdf

[17] Andrae Muys. 2008. A Concurrency Control Protocol for Multiversion RDF Datastores (Discussion Paper).

[18] Thomas Neumann and Gerhard Weikum. 2010. x-RDF-3X: Fast Querying, High Update Rates, and Consistency for RDF Databases. *Poceedings of the VLDB* 3 (Sept. 2010). DOI: http://dx.doi.org/10.14778/1920841.1920877

[19] John Ousterhout and Eric Stratmann. 2010. Managing State for Ajax-Driven Web Components. In *Proceedings of the 2010 USENIX Conference on Web Application Development (WebApps'10)*. Boston, MA, USA.

[20] Shirish Hemant Phatak and Badri Nath. 2004. Transaction-Centric Reconciliation in Disconnected Client-Server Databases. *Mobile Networks and Applications* (2004). DOI: http://dx.doi.org/10.1023/B:MONE.0000034700.03069.48

[21] Mahadev Satyanarayanan, James J. Kistler, Puneet Kumar, Maria E. Okasaki, Ellen H. Siegel, and David C. Steere. 1990. Coda: A Highly Available File System for a Distributed Workstation Environment. *IEEE Trans. Comput.* 39 (1990). Issue 4. DOI: http://dx.doi.org/10.1109/12.54838

[22] The W3C SPARQL Working Group. 2013. *SPARQL 1.1 Overview*. Recommendation. W3C. https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/

Decentralized Collaborative Knowledge Management using Git

Extended Abstract

Natanael Arndt
Agile Knowledge Engineering and Semantic Web (AKSW)
Institut fÃijr Informatik, Leipzig University
Institut fÃijr Angewandte Informatik (InfAI) e.V.
Leipzig, Germany
arndt@informatik.uni-leipzig.de

Michael Martin
Agile Knowledge Engineering and Semantic Web (AKSW)
Institut fÃijr Angewandte Informatik (InfAI) e.V.
Leipzig, Germany
martin@informatik.uni-leipzig.de

ABSTRACT

Apart from documents, datasets are gaining more attention on the World Wide Web. An increasing number of the datasets on the Web are available as Linked Data, also called the *Linked Open Data Cloud*[1] or *Giant Global Graph*[2]. Collaboration of people and machines is a major aspect of the World Wide Web and as well of the Semantic Web. Currently, the access to RDF data on the Semantic Web is possible by applying the Linked Data principles[3], and the SPARQL specification[4], which enables clients to access and retrieve data stored and published via SPARQL endpoints. RDF resources in the Semantic Web are interconnected and often correspond to previously created vocabularies and patterns. This way of reusing existing knowledge facilitates the modeling and representation of information and may optimally reduce the development costs of a knowledge base. As a result of the collaborative reuse process, structural and content interferences as well as varying models and contradictory statements are inevitable.

CCS CONCEPTS

• **Information systems → Query languages**; **Version management**; **Resource Description Framework (RDF)**; *Data management systems*; *Middleware for databases*; RESTful web services.

KEYWORDS

Versioning, RDF, Semantic Web, Git, Distributed Version Control System, Distributed Collaboration, Knowledge Engineering, Quit Store

ACM Reference Format:
Natanael Arndt and Michael Martin. 2019. Decentralized Collaborative Knowledge Management using Git: Extended Abstract. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316523

[1] http://lod-cloud.net/
[2] http://dig.csail.mit.edu/breadcrumbs/node/215
[3] http://www.w3.org/DesignIssues/LinkedData.html
[4] https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316523

Projects from a number of domains are striving for distributed models to collaborate on common knowledge bases. In the domain of e-humanities projects often come with a need to explore and track provenance and the evolution of the domain data [6, 7]. In the context of managing historical prosopographical data, the source of the statements is relevant to evaluate their credibility and to consider the influence of their environment. In libraries, metadata of electronic library resources are gathered and shared among stakeholders to collaboratively curate and manage the resources as Linked Data [2, 5]. In a collaborative data curation setup the origin of any statement needs to be identified in order to be able to track back the conclusion of license contracts and identify sources of defective metadata. But even enterprises have a need to manage data in distributed setups to organize the communication of data along supply chains or business processes [4].

Distributed systems such as the *Solid*[5] platform as an advancement of the architecture of a distributed semantic social network provide possibilities to collaborate in a distributed network. Nevertheless, the subject of collaboration is currently kept in a central place where all contributions are incorporated; the organization of a fully decentralized collaboration process is still subject to future work. In general, currently the collaboration on Linked Data Sets is mainly done by keeping a central version of a dataset. The systems available to collaborate on Linked Data are central SPARQL endpoints and Wiki systems where collaboration happens on a single, shared instance. This central approach for a synchronized state has drawbacks in scenarios in which the existence of different versions of the dataset is preferable. Furthermore, the evolution of a dataset in a distributed setup is not necessarily happening in a linear manner. Multiple versions of a dataset occur if the participants do not all have simultaneous access to the central dataset. If a consensus on the statements in a dataset is not yet reached, multiple viewpoints need to be expressed as different versions of the dataset. Hence, a system that fosters the evolution of a dataset in a distributed collaboration setup needs to **support divergence** of datasets as asynchrony and dissent; **reconcile diverged states** of datasets; and **synchronize** different distributed derivatives of the dataset. As a consequence of the reconciliation we also needs to identify possible occurring conflicts and contradictions, and offer workflows to resolve identified conflicts and contradictions. The dimensions of *consensus* vs. *dissent* and *synchronicity* vs. *asynchrony* are depicted in fig. 1. While the *dissent*-dimension comes with the

[5] https://solid.mit.edu/

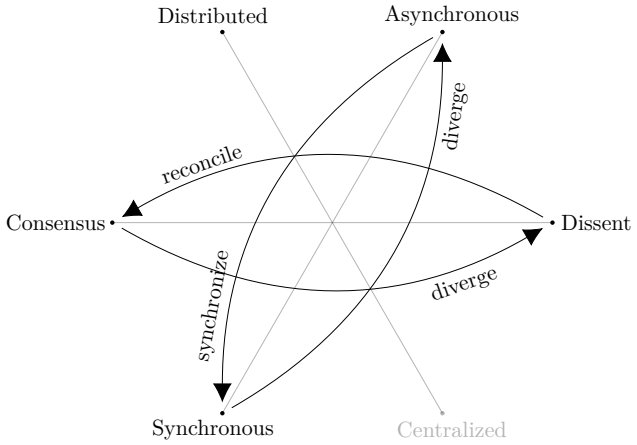


Figure 1: Collaboration can be organized *centralized* or *distributed*. When dealing with distributed collaboration, workspaces can *diverge* and the aspects *dissent* and *asynchrony* have to be considered. Even so supporting distribution, dissent, and asynchrony increases the flexibility of a collaboration process, collaboration aims for *consensus* which requires all participants to have access to a common shared knowledge, expressed in *synchronous* workspaces. Thus processes to *synchronize* and *reconcile* are needed.

collaborative character, *asynchrony* is introduced due to the *distributed* conception of our setup. Both of the dimensions can lead to a diverged state of a dataset in a collaborative curation scenario.

In the early days of computers, the term *software crisis* was coined to describe the immaturity of the software engineering process and software engineering domain. The process of creating software could be made more reliable and controllable by introducing software engineering methods. Version control is an important aspect to organize the collaborative evolution of software. Early *version control systems* (VCS), such as *CVS* and *Subversion*, allowed central repositories with a linear version history to be created. Distributed VCS (DVCS), such as *Darcs*, *Mercurial*, and *Git*, were developed to allow every member of a distributed team to fork the current state of the programs source code and individually contribute new features or bug-fixes as pull-requests. Learning from software engineering history where DVCS have helped to overcome the software crisis, we claim that adapting DVCS to Linked Data is a means to support decentralized and distributed collaboration processes in knowledge management. The subject of collaboration in the context of Linked Data are datasets instead of source code files. Similar to source code development with DVCS, individual local versions of a dataset are curated by data scientists and domain experts.

In our previously published paper [1] we present *Quit Store*, it was inspired by and it builds upon the successful Git system. The approach is based on a formal expression of evolution and reconciliation of distributed datasets. It provides support to branch, merge, and synchronize distributed RDF datasets. During the collaborative curation process, the system automatically versions the RDF dataset and tracks provenance information. The provenance information is expressed in RDF using PROV-O and can be accessed through a dedicated SPARQL 1.1 endpoint. To version the data, the system relies on the pure RDF data model and not on support for additional semantics such as OWL or SKOS. To support distributed collaboration we propose a methodology of using a Git repository to store the data in combination with a SPARQL 1.1 interface to access it. The SPARQL 1.1 interface provides an integration layer to make the collaboration features of Quit accessible to applications operating on RDF datasets. Most recently we have extended the Quit system with the Quit Editor Interface Concurrency Control [3] to support editors in managing overlapping operations. To reconcile diverged datasets a merge process is provided. The merge process is guarded by the specific merge strategies for RDF data: *Union Merge*, *All Ours/All Theirs*, *Three-Way-Merge*, and *Context Merge*. This setup can enable complex distributed collaboration strategies. As there is a big ecosystem of methodologies and tools around Git to support the software development process, the Quit Store can support the creation of such an ecosystem for RDF dataset management.

ACKNOWLEDGMENTS

This work was partly supported by a grant from the German Federal Ministry of Education and Research (BMBF) for the LEDS Project under grant agreement No 03WKCG11C the Federal Ministry for Economic Affairs and Energy (BMWi) for the Platona-M project under the grant number 01MT19005A, and the DFG project *Professorial Career Patterns of the Early Modern History: Development of a scientific method for research on online available and distributed research databases of academic history* under the grant agreement No 317044652.

REFERENCES

[1] Natanael Arndt, Patrick Naumann, Norman Radtke, Michael Martin, and Edgard Marx. 2018. Decentralized Collaborative Knowledge Management using Git. *Journal of Web Semantics* (2018). https://doi.org/10.1016/j.websem.2018.08.002

[2] Natanael Arndt, Sebastian Nuck, Andreas Nareike, Norman Radtke, Leander Seige, and Thomas Riechert. 2014. AMSL: Creating a Linked Data Infrastructure for Managing Electronic Resources in Libraries. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track*. Riva del Garda, Italy. http://ceur-ws.org/Vol-1272/paper_66.pdf

[3] Natanael Arndt and Norman Radtke. 2019. Conflict Detection, Avoidance, and Resolution in a Non-Linear RDF Version Control System: The Quit Editor Interface Concurrency Control. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion)*. San Francisco, CA, USA. https://doi.org/10.1145/3308560.3316519

[4] Marvin Frommhold, Natanael Arndt, Sebastian Tramp, and Niklas Petersen. 2016. Publish and Subscribe for RDF in Enterprise Value Networks. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 25th International World Wide Web Conference (WWW 2016)*. http://ceur-ws.org/Vol-1593/article-09.pdf

[5] Andreas Nareike, Natanael Arndt, Norman Radtke, Sebastian Nuck, Leander Seige, and Thomas Riechert. 2014. AMSL: Managing Electronic Resources for Libraries Based on Semantic Web. In *Proceedings of the INFORMATIK 2014: Big Data – Komplexität meistern*. Gesellschaft fÃijr Informatik e.V., Stuttgart, Germany. https://dl.gi.de/bitstream/handle/20.500.12116/2713/1017.pdf

[6] Thomas Riechert and Francesco Beretta. 2016. Collaborative Research on Academic History using Linked Open Data: A Proposal for the Heloise Common Research Model. *CIAN-Revista de Historia de las Universidades* 19, 0 (2016). http://e-revistas.uc3m.es/index.php/CIAN/article/view/3147

[7] Thomas Riechert, Ulf Morgenstern, Sören Auer, Sebastian Tramp, and Michael Martin. 2010. Knowledge Engineering for Historians on the Example of the Catalogus Professorum Lipsiensis. In *Proceedings of the 9th International Semantic Web Conference (ISWC2010)*. Springer, Shanghai, China. https://doi.org/10.1007/978-3-642-17749-1_15

Modelling the Provenance of Linked Data Interlinks for the Library Domain

Lucy McKenna
ADAPT Centre,
Trinity College Dublin,
Ireland
lucy.mckenna@adaptcentre.ie

Christophe Debruyne
ADAPT Centre,
Trinity College Dublin,
Ireland
christophe.debruyne@adaptcentre.ie

Declan O'Sullivan
ADAPT Centre,
Trinity College Dublin,
Ireland
declan.osullivan@cs.tcd.ie

ABSTRACT

As the Web of Data grows, so does the need to establish the quality and trustworthiness of its contents. Increasing numbers of libraries are publishing their metadata as Linked Data (LD). As these institutions are considered authoritative sources of information, it is likely that library LD will be treated with increased credibility over data published by other sources. However, in order to establish this trust, the provenance of library LD must be provided.

In 2018 we conducted a survey which explored the position of Information Professionals (IPs), such as librarians, archivists and cataloguers, with regards to LD. Results indicated that IPs find the process of LD interlinking to be a particularly challenging. In order to publish authoritative interlinks, provenance data for the description and justification of the links is required. As such, the goal of this research is to provide a provenance model for the LD interlinking process that meets the requirements of library metadata standards. Many current LD technologies are not accessible to non-technical experts or attuned to the needs of the library domain. By designing a model specifically for libraries, with input from IPs, we aim to facilitate this domain in the process of creating interlink provenance data.

CCS CONCEPTS

• **General and reference** → **Design**; • **Information systems** → **Data provenance**; **Digital libraries and archives**; **Semantic web description languages**; • **Human-centered computing** → *User centered design.*

KEYWORDS

linked data, semantic web, interlinking, provenance, library

ACM Reference Format:
Lucy McKenna, Christophe Debruyne, and Declan O'Sullivan. 2019. Modelling the Provenance of Linked Data Interlinks for the Library Domain. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316518

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316518

1 INTRODUCTION

The Semantic Web (SW) is an extension of the current Web where data is given well defined meaning and where the relationships between data, and not just documents, are defined in a common machine-readable format - creating a Web of Data [3]. Linked Data (LD) describes a set of best practices for publishing and interlinking this data on the SW, as per the principles defined by the W3C [2, 5]. These principles include the use of HTTP Uniform Resource Identifiers (URIs) as names for entities, such as works, people, places, and events, and also for retrieving data using the existing HTTP stack. A LD dataset is structured information encoded using the Resource Description Framework (RDF), the recommended model for representing and exchanging LD on the Web [30]. RDF statements take the form of subject-predicate-object triples, which can be organised in graphs. RDF requires that URIs are used to identify subjects and predicates - allowing for the resulting data to be understood by computers.

The SW and LD have the potential to transform the Web into a globally interlinked and searchable database rather than a disparate collection of documents [31]. This would allow for easier data querying and processing, and for the development of novel applications built on top of this Web of Data. With the Web being one of the first places where people search for information, one domain that is set to benefit from publishing to the SW are libraries. By using LD, libraries could improve the discoverability, searchability and interoperability of their data [13], which in turn would increase the use of their resources. However, since any individual can publish to the SW, it is crucial that libraries not only publish their metadata, but also the provenance of this metadata to the SW. Provision of provenance information allows potential users to establish the origin and trustworthiness of the data it describes. Given that libraries are considered authoritative sources of information, data from this domain is likely to be treated with increased credibility[25]. As such, if given access to descriptive provenance data, Web users are likely to engage with library LD datasets with increased confidence and frequency.

Though the number of libraries publishing to the SW is growing, uptake is still relatively slow due to the range of challenges faced by these institutions when using LD, including a lack guidelines, financial constraints, data quality concerns, URI maintenance issues, and software complexity [17, 22, 29]. A 2018 survey explored the position of 185 Information Professionals' (IPs) with regards to LD and results highlighted LD interlinking as a task that IPs find to be particularly challenging [21]. In response to this, we developed a LD interlinking approach for the library domain called NAISC - the Novel Authoritative Interlinking of Schema and Concepts. The

aim of NAISC is to improve LD interlinking accessibility to non-technical experts. This being achieved the the iterative design and user-testing of a graphical user-interface (GUI) which guides users through a step-by-step process of interlink generation[1]. NAISC specifically targets IPs by providing access to datasets and ontologies commonly used in the library domain. NAISC also reduces the need for expert LD knowledge by suggesting suitable link-types based links created by the user using natural-language relationship terms.

With one of the fundamental prerequisites of the SW being the existence of large amounts of meaningfully interlinked resources [5], it is not only important that libraries are facilitated to interlink their data to a range of authoritative sources, but that the trustworthiness of such interlinks is established through the provision of provenance data. Thus, as part of NAISC, we developed a provenance model for the LD interlinking process that meets the unique requirements of the library domain.

Our paper describes the NAISC provenance model, and it is structured as follows: a Background section provides information on LD Interlinking and LD Provenance; the Aims and Provenance Requirements for our model are then discussed; this is followed by a description of our Provenance Model with a Demonstration of potential uses. Lastly, the Future Directions and Conclusion of our research are discussed.

2 BACKGROUND

In the following section LD Interlinking and LD Provenance are defined and discussed in the context of our research within the library domain.

2.1 Linked Data Interlinking

Data linking describes the task of determining whether a URI, used to identify an entity, can be linked to another URI as a way of representing that they both describe the same Thing or as a way of indicating that they are related in some capacity [10]. LD interlinks are known as *typed links*, so called because the linking property, or predicate, describes the type of relationship between the subject URI and the object URI [26]. The property used to describe the relationship between two URIs is known as a link-type. In the context of our research, LD interlinking specifically refers to the process of creating an interlink between two URIs from different data sources.

2.2 Linked Data Provenance

Provenance data provides information on the people, institutions, resources, and processes involved in creating a piece of data [24]. This data can be used in order to ascertain whether information is trustworthy and as a means of determining data quality [18, 20]. Since any individual or group can publish to the SW, it is crucial that libraries publish the provenance of their interlinks as this would allow researchers to establish the origin of the data. Given that libraries are considered authoritative sources of information [25], it is possible that interlinks from this domain will be deemed trustworthy and thus used more frequently. In the context of our research, interlinks with rich data provenance are considered authoritative LD interlinks.

There are a number of provenance models that have been developed for use with LD including the Provenance Vocabulary [16], the Open Provenance Model (OPM) [23], Provenance Authoring and Versioning ontology (PAV) [9], Provenir [28], and the W3C recommended standard, PROV Ontology (PROV-O) [19]. The PROV Data Model, shown in Figure 1, is a Web Oriented provenance standard, developed by the W3C Provenance Working Group [19], for the representation and exchange of provenance information [24]. The model can be used to describe the Entities (physical, digital or conceptual object), Agents (person, organisation, software) and Activities involved the process of creating a specific Entity [19].

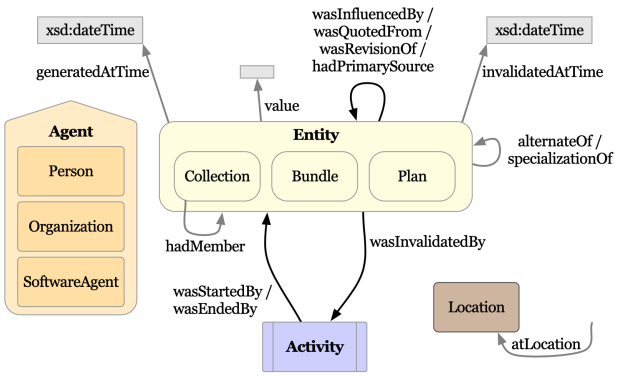


Figure 1: PROV Data Model

2.2.1 Provenance of Digital Resources. The Open Archival Information System (OAIS) [11] and Preservation Metadata: Implementation Strategies (PREMIS) [27] are widely accepted standards for digital preservation. Both OAIS and PREMIS require the provision of provenance information when archiving digital resources so as to maintain their long-term use and preservation. In the library domain, data provenance requires the inclusion information on where, when, by whom and how a resource was created [20]. Given that data provenance is likely to play an important role in establishing the trustworthiness of LD, it seems appropriate that these provenance standards should also be applied to the creation of interlinks. However, LD software typically only provides provenance information on resource ownership, as well as time-stamps for resource creation or modification [14]. As such, there is a need for a LD provenance model that captures the rich data required by the library domain in order to create authoritative interlinks.

3 NAISC PROVENANCE MODEL

As mentioned, we are currently developing an interlinking approach specifically for the library domain called NAISC. A component of the interlinking approach is the NAISC Provenance Model which provides authoritative origin data for the interlinks generated as part of the larger NAISC approach. The design of the provenance model is discussed below.

[1]See https://www.scss.tcd.ie/~mckennl3/naisc for a video demonstration of NAISC.

3.1 Provenance Requirements

A set of user requirements for the provenance model were distilled from the results of our international survey of 185 IPs [21]. The majority of participants (56%) came from an Academic Library perspective, thus the results of the survey are most applicable to this domain. Additionally most participants had some prior knowledge of the SW (84%) and LD (90%). The provenance requirements included:

- Allow for different levels of granularity e.g. view provenance for a set of links and for an individual link.
- Keep track of revisions made to interlinks.
- Link to the dataset/source of a subject or object entity.
- Link to the creator of the interlinks and the provenance data.
- Allow for the justification of linking a pair of entities.
- Allow for the justification of the link-type.

Further requirements for the provenance model were established from a series of ontological competency questions [4, 15], see Table 1. These questions were inspired by common requirements for data provenance on the SW [14].

Table 1: Interlink Provenance Competency Questions

Who created the link?	How can the dataset be accessed?
How was the link created?	Who published the dataset?
Why was the link created?	When was the link modified?
Where was the link created?	Who modified the link?
When was the link created?	How was the link modified?
What resources are linked?	Why was the link modified?
Why was the link created?	Who created the link provenance?
What datasets are linked?	When was the provenance created?

3.2 PROV-O Extension

PROV-O was used as the foundation of our interlink provenance model as it is a W3C recommended standard [19]. It also provides a model for general provenance descriptions which can then be extended for the needs of domain specific purposes [9]. Existing PROV-O classes, sub-classes and properties were used to describe the who (prov:Person), where (prov:Organisation) and when (prov:generatedAtTime) interlinks were created. We then extended PROV-O, see Figure 2, in order to add interlink specific sub-classes and properties. This extension, called NaiscProv, describes how (naiscProv:InterlinkCreationActivity) and why (naiscProv:hasJustification) interlinks were created.

Dublin Core (dcterms) [6] and FOAF [7] ontologies were used to provide richer descriptions of subject and object entities. The VoID Vocabulary [1] was also used in order to describe the datasets, or sources, of these entities.

3.3 Graph Structure

Our Provenance Model, as seen in Figure 3, incorporates three graphs:

(1) Interlink Graph - a named graph containing a set of interlinks. A named graph is a sub-graph that contains a set of triples and that has been assigned a unique name in the form of a URI [8]. Named graphs allow collections of triples to be published as independent units. Named graphs are often used in the process of provenance data generation as they allow for the assertion of statements relating to a specific set of triples in a dataset [12]. In the case of NAISC, an Interlink Graph will contain a set of finalised interlinks associated with a particular dataset, or part of a dataset, that are ready to be published to the SW. Over time, as interlinks are added, modified or deleted from the dataset, new Interlink Graphs will be created. Having these multiple Interlink Graphs allows for the provenance of individual interlinks to be maintained over time in a more simplified manner.

(2) Provenance Graph - a prov:Bundle containing the origin data of the interlinks in an Interlink Graph. In the PROV Data Model, a Bundle is a named set of provenance descriptions that can be used to describe the creation and modification of an entity or group of entities [19]. As a Bundle is itself an entity, the provenance of the Provenance Graph can also be captured. Every Interlink Graph will have a corresponding Provenance Graph which captures who, where, when, why and how the interlinks contained in the graph were created.

(3) Relationship Graph - represents the relationship between an Interlink Graph and a Provenance Graph (prov:hasProvenance).

The purpose of these graphs is to allow the user to explore the different sets of interlinks, and also to explore the provenance information for the interlinks. The Interlink Graph allows the user

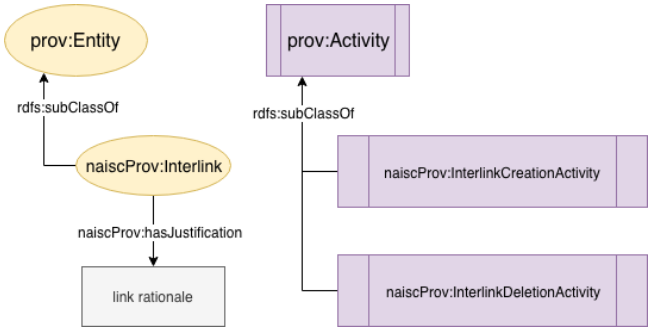


Figure 2: NaiscProv PROV-O Extension

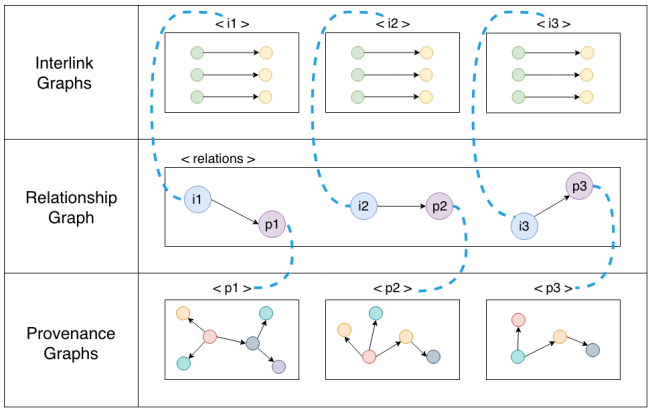


Figure 3: NAISC Provenance Graph Structure

to view a particular set of LD interlinks created using the NAISC Approach. Should the user wish to review the provenance of this set of interlinks, the Relationship Graph can be used to direct the user to the associated Provenance Graph associated with the Interlink Graph. The Provenance Graph provides origin information for each of the interlinks as well as for the interlink creation, revision and deletion processes.

Separating the data in this manner simplifies some of the queries that users could formulate and run over the data whilst still allowing for queries that span across graphs, as facilitated by the relationship layer.

4 DEMONSTRATION

In the following section a simple provenance graph for the creation, revision and deletion of an interlink shall be demonstrated using the NAISC Provenance Model.

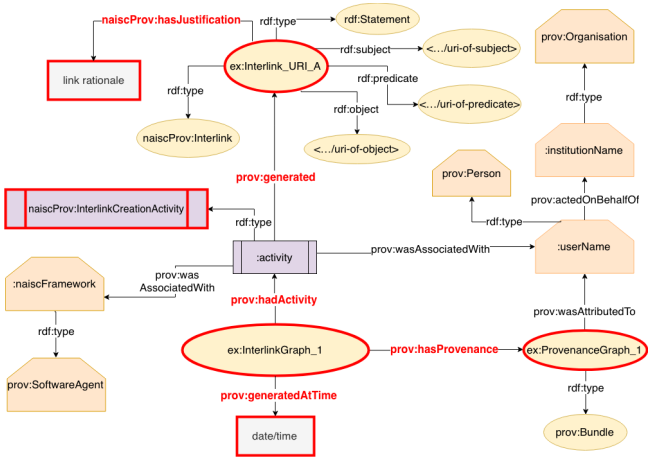


Figure 4: Interlink Creation Provenance Graph

4.1 Interlink Creation

Figure 4 demonstrates a snippet of a Provenance Graph describing the process of creating an interlink. Here an Interlink Graph (ex:InterlinkGraph_1), containing a set of interlinks, is stated to have had a creation activity (naiscProv:InterlinkCreationActivity) which generated a specific interlink (ex:Interlink_URI_A). The interlink is given a Unique Resource Identifier (URI) and its subject, predicate and object are described using RDF Reification. RDF Reification allows for each interlink to be given its own URI, and also allows for use of the naiscProv:hasJustification property. This property provides an opportunity for the inclusion of rationale, or 'why' provenance, data. Information that could be captured here includes justifying why the subject and predicate entities were interlinked, as well as the choice of predicate.

Other important provenance information included in the Provenance Graph is the generation date/time of the Interlink Graph (prov:generatedAtTime), as well as the Agents responsible for creating the link such as NAISC (prov:SoftwareAgent), the cataloguer (prov:Person), and the cataloguer's institution (prov:actedOnBehalfOf). Also included is the provenance of the Provenance Graph itself

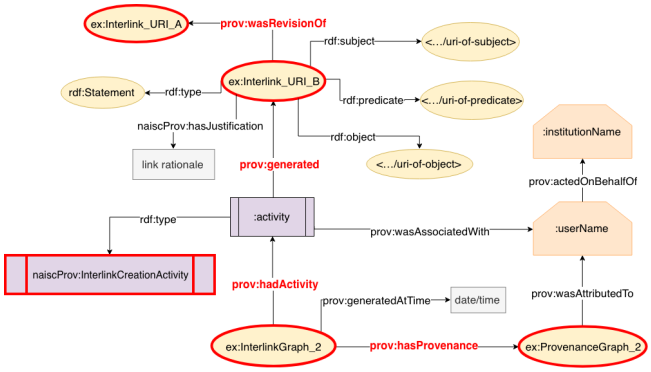


Figure 5: Interlink Revision Provenance Graph

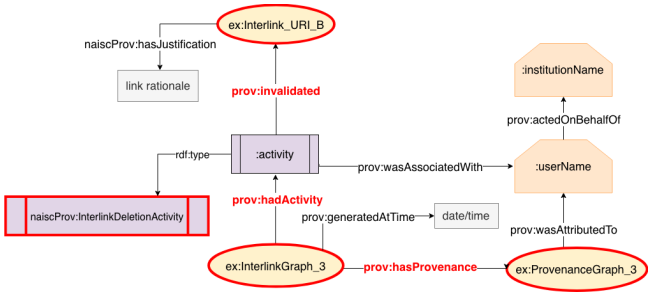


Figure 6: Interlink Deletion Provenance Graph

(prov:hasProvenance). Other information that could be added to the graph includes the source of the subject and/or object entity (dcterms:isPartOf, void:Dataset).

4.2 Interlink Revision

Figure 5 demonstrates a Provenance Graph describing the revision of an interlink using the NAISC Provenance Model. Here a new Interlink Graph (ex:InterlinkGraph_2) is stated to have had a creation activity which generated a new Interlink (ex:Interlink_URI_B). The property prov:wasRevisionOf is used to describe how the new interlink (ex:Interlink_URI_B) is a modified version of (ex:Interlink_URI_A). The rationale for modifying the interlink can be provided through the use of the naiscProv:hasJustification property.

4.3 Interlink Deletion

Figure 6 demonstrates the deletion of an interlink using the NAISC Provenance Model. In this instance a new Interlink Graph (ex:InterlinkGraph_3) is stated to have had a deletion activity (naiscProv:InterlinkDeletionActivity). This activity resulted in the invalidation, or deletion, (prov:inalidated) of ex:Interlink_URI_B. The reason for deleting the interlink can be provided using the naiscProv:hasJustification property. As seen previously, the relationship between the new Interlink Graph (ex:Interlin-kGraph_3), and its predecessor (ex:InterlinkGraph_2) is conveyed using prov:wasRevisionOf.

5 CONCLUSION

One of the main benefits of the SW is the ability to interlink related entities across datasets. However, such interlinks can only be meaningfully used if their origin and creation processes are exposed to users. This data enables the assessment of the context in which the interlink was created as well as its quality and validity. As part of our research we developed an interlink provenance model specifically for the library domain. This provenance model meets the unique requirements of IPs, as discussed in Section 3.1, particularly that of the provision of 'why' provenance information which is not catered for in other provenance models. This provenance data will add to the trustworthiness of library LD which in turn may increase the use of interlinks published as part of these datasets. By incorporating our Provenance Model into the NAISC framework, IPs will be able to easily create richer provenance data for the interlinks they generate in a using NAISC's GUI.

6 FUTURE DIRECTIONS

As mentioned, the NAISC Provenance Model forms part of the NAISC Approach to LD Interlinking. An accompanying graphical user interface has been developed as a means of guiding users through the steps of LD interlink creation and provenance generation, as proposed by our approach. The NAISC Provenance Model will be evaluated as part of the user-testing of the NAISC Approach. This will involve assessing the efficacy of the model in capturing the provenance of interlinks as well as the perceived influence of the provenance data on the trustworthiness of the interlinks.

ACKNOWLEDGMENTS

This study is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre for Digital Content Platform Research (http://www.adaptcentre.ie/) at Trinity College Dublin.

REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. 2011. Describing Linked Datasets with the VoID Vocabulary. Retrieved November 2018 from https://www.w3.org/TR/void/.
[2] T. Berners-Lee. 2006. Linked Data. Retrieved March 2019 from https://www.w3.org/DesignIssues/LinkedData.
[3] T. Berners-Lee, J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American* 284, 5 (2001), 1–5.
[4] C. Bezerra, F. Freitas, and F. Santana. 2013. Evaluating ontologies with competency questions. In *In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 3. 284–285.
[5] C. Bizer, T. Heath, and T. Berners-Lee. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems* 5, 3 (2009), 1–22.
[6] Dublin Core Metadata Initiative Usage Board. 2014. Dublin Core Metadata Initiative Metadata Terms. Retrieved November 2018 from http://dublincore.org/documents/2012/06/14/dcmi-terms/.
[7] D. Brickley and L. Miller. 2014. FOAF Vocabulary Specification 0.99. Retrieved November 2018 from http://xmlns.com/foaf/spec/.
[8] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. 2005. Named graphs. *Web Semantics: Science, Services and Agents on the World Wide Web* 3, 4 (2005), 247–267.
[9] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark. 2013. PAV ontology: provenance, authoring and versioning. *Journal of Biomedical Semantics* 4, 1 (2013), 37.
[10] A. Ferrara, A. Nikolov, and F.çois Scharffe. 2011. Data linking for the semantic web. *International Journal on Semantic Web and Information Systems (IJSWIS)* 7, 3 (2011), 46–76.
[11] Consultative Committee for Space Data Systems. 2002. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1, 1-1. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.
[12] T. Gibson, K. Schuchardt, and E. Stephan. 2009. Application of named graphs towards custom provenance views. In *Paper presented at the First workshop on on Theory and practice of provenance, San Francisco, CA.*
[13] B.M. Gonzales. 2014. Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record. *Information Technology and Libraries* 33, 4 (2014), 10–22.
[14] P. Groth, Y. Gil, J. Cheney, and S. Miles. 2012. Requirements for provenance on the web. International Journal of Digital Curation. *nternational Journal of Digital Curation* 7, 1 (2012), 39–56.
[15] M. Gruninger and M.S. Fox. 1995. Methodology for the Design and Evaluation of Ontologies. In *In Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing.*
[16] O. Hartig and J. Zhao. 2010. Publishing and Consuming Provenance Metadata on the Web of Linked Data.. In *Paper presented at the IPAW 2010, Berlin, Heidelberg.*
[17] R. Hastings. 2015. Linked Data in Libraries: Status and Future Direction. *Computers in Libraries* 35, 9 (2015), 12–16.
[18] S. Kumar, M. Ujjal, and B. Utpal. 2013. Exposing MARC 21 Format for Bibliographic Data As Linked Data With Provenance. *Journal of Library Metadata* 13, 2-3 (2013), 212–229.
[19] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, and D. et al Corsar. 2013. Prov-o: The prov ontology. W3C Recommendation. World Wide Web Consortium. Retrieved November 2018 from https://www.w3.org/TR/prov-o/.
[20] C. Li and S. Sugimoto. 2014. Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata. In *Proceedings of the IInternational Conference on Dublin Core and Metadata Applications, Sao Paulo, Brazil.*
[21] L. McKenna, C. Debruyne, and D. O'Sullivan. 2018. Understanding the Position of Information Professionals with regards to Linked Data: A Survey of Libraries, Archives and Museums. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries.* 7–16.
[22] E.T. Mitchell. 2016. Library Linked Data: Early Activity and Development. *Library Technology Reports* 52, 1 (2016), 5–33.
[23] L. Moreau, B. Clifford, J. Freire, J. Futrelle, and Y. et al Gil. 2011. The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems* 27, 6 (2011), 743–756.
[24] L. Moreau, P. Groth, J. Cheney, T. Lebo, and S. Miles. 2015. The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web* 35 (2015), 235–257.
[25] P. Neish. 2015. Linked data: what is it and why should you care? *The Australian Library Journal* 64, 1 (2015), 3–10.
[26] Georg Neubauer. 2017. Visualization of typed links in Linked Data. *Mitteilungen der Vereinigung ÃŰsterreichischer Bibliothekarinnen und Bibliothekare* 70 (09 2017), 179. https://doi.org/10.31263/voebm.v70i2.1748
[27] Library of Congress. 2018. PREMIS: Preservation Metadata Maintenance Activity. Retrieved January 2019 from http://www.loc.gov/standards/premis/.
[28] S. S. Sahoo and A. P. Sheth. 2009. Provenir ontology: Towards a Framework for eScience Provenance Management. In *Microsoft eScience Workshop, Pittsburgh, PA Oct 15-17.*
[29] K. Smith-Yoshimura. 2018. Analysis of 2018 international linked data survey for implementers. *Code4Lib* 42 (2018).
[30] W3C. 2014. RDF: Resource Description Framework. https://www.w3.org/RDF/.
[31] W3C. 2015. Semantic Web. https://www.w3.org/standards/semanticweb/.

TISCO: Temporal Scoping of Facts

Anisa Rula*
University of Milano-Bicocca
Milan, Italy
anisa.rula@unimib.it

Matteo Palmonari
University of Milano-Bicocca
Milan, Italy
matteo.palmonari@unimib.it

Simone Rubinacci
University of Oxford
Oxford, United Kingdom
s.rubinacci@stats.ox.ac.uk

Axel-Cyrille Ngonga Ngomo
University of Paderborn
Paderborn, Germany
axel.ngonga@upb.de

Jens Lehmann
University of Bonn
Bonn, Germany
lehmann@cs.uni-bonn.de

Diego Esteves
University of Bonn
Bonn, Germany
esteves@cs.uni-bonn.de

Andrea Maurino
University of Milano-Bicocca
Milan, Italy
andrea.maurino@unimib.it

ABSTRACT

Some facts in the Web of Data are only valid within a certain time interval. However, most of the knowledge bases available on the Web of Data do not provide temporal information explicitly. Hence, the relationship between facts and time intervals is often lost. A few solutions are proposed in this field. Most of them are concentrated more in extracting facts with time intervals rather than trying to map facts with time intervals. This paper studies the problem of *determining the temporal scopes of facts*, that is, deciding the time intervals in which the fact is valid. We propose a generic approach which addresses this problem by curating temporal information of facts in the knowledge bases. Our proposed framework, Temporal Information Scoping (TISCO) exploits evidence collected from the Web of Data and the Web. The evidence is combined within a three-step approach which comprises matching, selection and merging. This is the first work employing matching methods that consider both a single fact or a group of facts at a time. We evaluate our approach against a corpus of facts as input and different parameter settings for the underlying algorithms. Our results suggest that we can detect temporal information for facts from DBpedia with an f-measure of up to 80%.

CCS CONCEPTS

• **Information systems** → **Resource Description Framework (RDF)**; *Incomplete data*; • **Computing methodologies** → *Information extraction.*

KEYWORDS

Temporal Information Extraction; Temporal Semantic Web; Temporal Scoping; Fact Checking

*Principal corresponding author.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316524

ACM Reference Format:
Anisa Rula, Matteo Palmonari, Simone Rubinacci, Axel-Cyrille Ngonga Ngomo, Jens Lehmann, Diego Esteves, and Andrea Maurino. 2019. TISCO: Temporal Scoping of Facts. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316524

1 INTRODUCTION

The Web of Data can be regarded as a dynamic environment where information can change rapidly and cannot be assumed to be static [6]. Changes in the Web of Data sources should reflect changes in the real world, otherwise data can soon become outdated. Some facts are not time variant and thus do not change over time e.g. `<CristianoRonaldo, bornIn, Portugal>` while others need to be checked for their *veracity* [3]. In this work, we focus on a particular case of veracity of facts that is the *validity time* which refers to the time interval in which the fact is true, e.g. `<CristianoRonaldo, playFor, ManchesterUnited>` refers to a fact valid from 2003 to 2009.

Most of knowledge bases store facts under a historical perspective. Facts in this knowledge bases have been true sometime until the current time. We refer to the representation of these facts in the knowledge bases as to *historification of dynamic facts*. Historification of dynamic facts, which is frequently found in many and prominent datasets in Linked Data (LD), fails to provide details about the time interval when the facts have been true. These knowledge bases adopt a *temporal flattening approach* of representing dynamic facts. The incompleteness and the inaccuracy of temporal information in LD [9] is often due to the information extraction process (that can be error prone) or to the representation model (that requires very sophisticated meta-modeling strategies to represent *versioning metadata* in RDF). For instance, in DBpedia it is not possible to know the time interval of the fact `<CristianoRonaldo, playFor, ManchesterUnited>` since all time points are associated directly with the entity rather than the fact and the semantics of the predicate is the same for the starting and the ending time points (e.g. `<CristianoRonaldo, year, 2009>`).

Despite the importance of the relationship between facts and time intervals, very few solutions are proposed. Most of them are concentrated more in extracting facts with time intervals from text [5, 8, 12] rather than trying to map facts with time intervals. More recent works [2, 7] provide an enrichment of time intervals in the knowledge bases. The system CoTS provided in [11] is similar to our system since it also detects validity time of facts. In contrast to our approach, CoTS relies on document metadata such as its creation date, to assign time intervals to facts. To the best of our knowledge, this is the first work employing both local and global matching approach, and a system for mapping facts to time intervals.

To map those facts to the correct time intervals, we have to address two main challenges. First, the set of time intervals is created as a combination of all time points extracted from the knowledge base for each entity where the starting time point is smaller than the ending time point. This set needs to be reduced since it contains also noisy intervals such as all intervals starting with the birth year. Second, to find the correct intervals, we need to extract supporting evidence from external sources that indicates how often a fact occurs with each time point, and subsequently predict the possible time interval of each fact based on the acquired evidence.

In this work, we focus on curating time intervals associated with facts. We introduce an approach for detecting the temporal scope of facts referred to by triples (short: the temporal scope of the triples). Given a fact (i.e., an RDF triple), our approach aims to detect the time points at which the temporal scope of the triple begins and ends. Two sources can be envisaged for gathering such information: the document Web and LD. Our approach is able to take advantage of both: the Web is made use of by extending upon a fact validation approach [4], which allows detecting Web documents which corroborate a triple. In contrast to typical search engines, the system does not just search for textual occurrences of parts of the statement, but tries to find webpages which contain the actual statement phrased in natural language. The second source of information for time scopes is the Web of Data itself. Here, we use DBpedia, for possible time scopes and devise an algorithm for combining the results extracted from Web documents with those fetched from RDF sources. The algorithm consists of three main steps: First, the evidence extracted from Web documents is *matched* against a set of relevant time intervals to obtain a significance score for each interval. Second, a small set of more significant intervals is *selected*. Finally, the selected intervals are *merged*, when possible, by considering their mutual temporal relations. The set of disconnected intervals [1] returned by the algorithm defines the temporal scope of the fact. We also propose two *normalization* strategies that can be applied to the data extracted from Web documents before running the algorithm, to account for the significance of dates appearing in the documents corroborating the input fact.

This article makes the following contributions: i) We present an approach for modelling a space of relevant time intervals for a fact starting from dates extracted from RDF triples. ii) We devise a three-phase algorithm for temporal scoping, i.e. for mapping facts to sets of time intervals, which integrates the previous steps via matching, selection and merging. iii) We describe two matching methods that consider facts in isolation or cluster them according to the main entity. iv) We provide TISCO, a running prototype; the first system able to provide temporally annotated facts which are modelled according to a relationship-centric perspective [9].

We provide an extension of [10] as: i) We describe in detail more alternative solutions, including an additional function in the matching phase of our approach and a normalization function of occurrences of dates; ii) We developed a prototype for annotating facts with temporal information and show how all the different matching functions and their combinations can be integrated in one framework; iii) We extend the experimental results showing the effectiveness of our results.

2 CONCLUSIONS

This paper studies the problem of determining and mapping time intervals to dynamic facts. We proposed a framework comprising several functions and configuration parameters that can efficiently provide the matching and the selection of the set of time intervals that maximize the effectiveness of our approach. In addition, we proposed a running prototype, TISCO that will support the users in exploring facts with temporal scopes and simplify the testing of new algorithms for matching and selection functions. We evaluated our approach on facts extracted from DBpedia by using cleaned-up temporal scopes extracted from Yago2. We provide the benchmarking dataset to support the users in testing their approaches since it is difficult to perform the task at hand. Our approach is also tested on google books dataset and it outperforms one of the state-of-the-art approaches applied on the same dataset.

REFERENCES

[1] James F Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM* 26, 11 (1983), 832–843.

[2] Melisachew Wudage Chekol, Giuseppe PirrÃš, Joerg Schoenfisch, and Heiner Stuckenschmidt. 2017. Marrying Uncertainty and Time in Knowledge Graphs.. In *AAAI*, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 88–94.

[3] Diego Esteves, Anisa Rula, Aniketh Janardhan Reddy, and Jens Lehmann. 2018. Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. *J. Data and Information Quality* 9, 3 (2018), 16:1–16:26.

[4] Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, and René Speck. 2015. DefactoâĂŤtemporal and multilingual deep fact validation. *Web Semantics: Science, Services and Agents on the World Wide Web* 35 (2015), 85–101.

[5] Dirk Hovy, James Fan, Alfio Gliozzo, Siddharth Patwardhan, and Chris Welty. 2012. When did that happen?: linking events and relations to timestamps. In *13th EACL*. 9.

[6] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. 2013. Observing Linked Data Dynamics. In *ESWC*. 213–227.

[7] Erdal Kuzey, Jannik Strötgen, Vinay Setty, and Gerhard Weikum. 2016. Temponym Tagging: Temporal Scopes for Textual Phrases. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. International World Wide Web Conferences Steering Committee, 841–842.

[8] Xiao Ling and Daniel S Weld. 2010. Temporal information extraction. In *25th AAAI*.

[9] Anisa Rula, Matteo Palmonari, Andreas Harth, Steffen Stadtmüller, and Andrea Maurino. 2012. On the Diversity and Availability of Temporal Information in Linked Open Data. In *11th ISWC*.

[10] Anisa Rula, Matteo Palmonari, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Jens Lehmann, and Lorenz Bühmann. 2014. Hybrid Acquisition of Temporal Scopes for RDF Data. In *11th ESWC*. 488–503. https://doi.org/10.1007/978-3-319-07443-6_33

[11] Partha Pratim Talukdar, Derry Tanti Wijaya, and Tom Mitchell. 2012. Coupled temporal scoping of relational facts. In *5th WSDM*. 73–82.

[12] Naushad UzZaman and James F Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *SemEval*. ACL, 276–283.

Summarizing Entity Temporal Evolution in Knowledge Graphs

Mayesha Tasnim*
RWTH Aachen
mayesha.tasnim@rwth-aachen.de

Diego Collarana
Fraunhofer IAIS & Univ. of Bonn
collaran@cs.uni-bonn.de

Damien Graux
Fraunhofer IAIS
damien.graux@iais.fraunhofer.de

Fabrizio Orlandi
ADAPT Centre, Trinity College Dublin
orlandif@tcd.ie

Maria-Esther Vidal
TIB, L3S & Leibniz Univ. of Hannover
maria.vidal@tib.eu

ABSTRACT

Knowledge graphs are dynamic in nature, new facts about an entity are added or removed over time. Therefore, multiple versions of the same knowledge graph exist, each of which represents a snapshot of the knowledge graph at some point in time. Entities within the knowledge graph undergo evolution as new facts are added or removed. The problem of automatically generating a summary out of different versions of a knowledge graph is a long-studied problem. However, most of the existing approaches are limited to a pairwise version comparison. This limitation makes it difficult to capture a complete evolution out of several versions of the same knowledge graph. To overcome this limitation, we envision an approach to create a summary graph capturing temporal evolution of entities across different versions of a knowledge graph. The entity summary graphs may then be used for documentation generation, profiling or visualization purposes. First, we take different temporal versions of a knowledge graph and convert them into RDF molecules. Secondly, we perform Formal Concept Analysis on these molecules to generate summary information. Finally, we apply a summary fusion policy in order to generate a compact summary graph which captures the evolution of entities.

KEYWORDS

RDF Knowledge Graph, RDF Molecules, Entity Evolution

ACM Reference Format:
Mayesha Tasnim, Diego Collarana, Damien Graux, Fabrizio Orlandi, and Maria-Esther Vidal. 2019. Summarizing Entity Temporal Evolution in Knowledge Graphs. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316521

1 INTRODUCTION

Knowledge graphs evolve over time with the addition of new entities and relationships, or the modification of its existing ones [4]. It is often the case that different versions of an RDF graph are maintained separately. For example, DBpedia releases a new version of its datasets on a yearly basis. Typically most applications based on knowledge graphs are concerned with the latest version available at any point in time. These graphs contain the most updated information, however, they are missing the knowledge about how these entities transform over multiple versions. From the perspective of a single entity, it can often be interesting to observe how it evolves over time. This information can be new knowledge that can add value to the existing knowledge graph. In this work, we propose a technique using RDF molecules and Formal Concept Analysis to summarize the temporal evolution of knowledge graphs.

Consider an RDF data entity representing the same person in different temporal versions of a knowledge graph, as depicted in Figure 1. The updates in the properties of this entity correspond to real-world changes over time. For example, the person might *inter alia* relocate to different cities or change job title. New properties may also be added to this entity, e.g. when this person becomes a parent for the first time. If only *some* versions of this knowledge graph are considered and others are ignored, the information obtained is valid solely in the context of that time period, and the knowledge about how this entity evolved over time is lost. For instance, considering the versions in Figure 1a and Figure 1d (years 2010 and 2016), it can be observed that this person has the same spouse *P*. But this observation misses the information that the person had a different spouse *Q* from 2012 to 2014 (see Figures 1b & 1c).

We motivate our work using this problem and propose a technique to summarize the evolution of entities in knowledge graphs. In this technique we identify entities in different versions of a knowledge graph that correspond to the same real-world object and apply formal concept analysis to generate a summary of the evolution of these entities. The generated *summary molecule* contains a compact representation of all the object and data properties the molecule was connected to over time, along with the temporal information indicating the ranges of validity for each property. In order to ensure compactness, the summary molecule contains each distinct object or data property only once, as shown in Figure 1e.

2 RELATED WORK

Knowledge graphs are becoming increasingly dynamic in nature and approaches have been proposed [9] to (i) detect changes during their evolution, (ii) represent change information (using vocabularies) and (iii) propagate changes to replicas or federated systems [2]. Approaches for change detection mainly focus on computing "deltas" (or changesets[1]) between two versions of a knowledge

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316521

[1] https://www.w3.org/2009/12/rdf-ws/papers/ws07 (accessed on 22/02/2019)

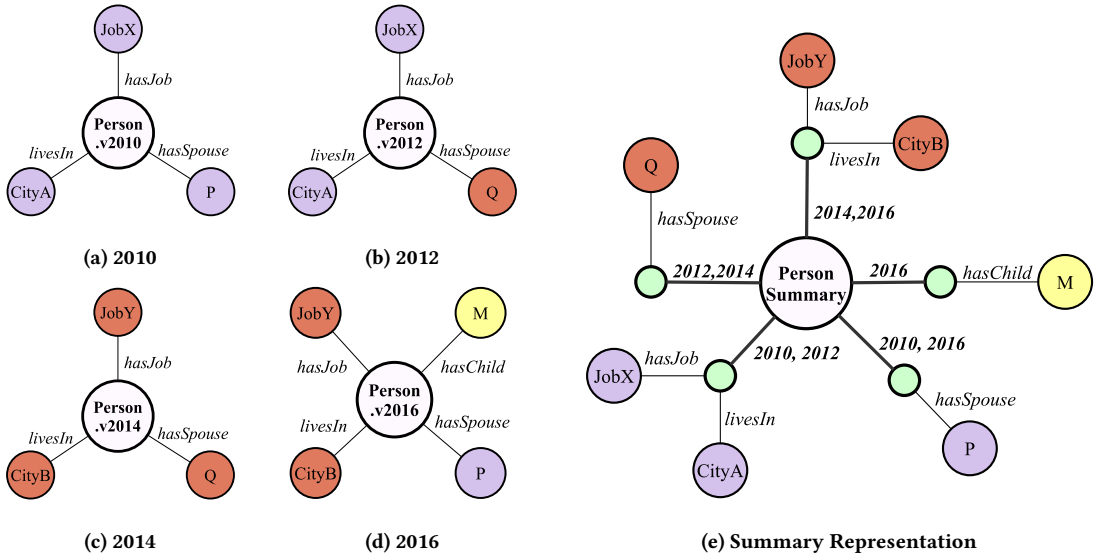


Figure 1: Visual representation of a person info obtained from four yearly molecules.

graph at different granularity levels [9]: dataset, resource and statement level. The Talis Changeset Vocabulary[2] defines a set of terms for describing changes to resource descriptions. In the context of this vocabulary, a resource description is the set of triples that includes a description of a resource. The DELTA-LD framework detects and classifies the changes between two versions of a linked dataset and represents them with the DELTA-LD change model [8]. DSNotify [6] is a change-detection system capable of detecting and fixing broken links between resources in two different versions of a dataset. In [7], in order to study the dynamics of LOD, the authors propose a framework for detecting and analysing changes and the evolution history of LOD datasets. Moreover, changes are automatically categorised according to their level of complexity and are represented using an ontology, hence allowing SPARQL to be used to query data changes. Specific queries need to be constructed in order to extract information about the history of changes for a particular class/entity across different versions. In contrast, our approach allows for automatic extraction and exploration of all changes of a class/entity over time.

Producing a summary of an entity evolution can be also tackled as an integration problem. Integration frameworks can be used to solve the problem of extracting the temporal evolution of an entity. In this regard, a central part of this work is represented by the MINTE approach [1]. We adopt the fusion policies for data integration described by Collarana et al. in order to produce a temporal evolution summary of entities across different versions of a knowledge graph. Similar to MINTE, ODCleanStore [5] is an ETL framework for integrating RDF data. It relies on SILK to perform instance matching and provides custom data fusion modules to merge the data of the discovered matches. The aim of the said integration frameworks is to map different data sources with possibly varying schema, i.e., they perform inter-schema mapping. However, even in these cases, creating a summarized view of an entity could only be supported on a superficial level by processing and filtering query results. Hence, we identify the need for an approach to automatically produce summaries of entities' evolution, as described in the following sections.

[2]http://vocab.org/changeset/schema (accessed on 22/02/2019)

3 THE APPROACH

3.1 Preliminaries

Given different versions of a knowledge graph, e.g., DBpedia 2010, 2012, 2014, and 2016, and an entity type, e.g., Person. Our approach automatically produces a summary of evolution over time of the entities under the specified type. Each entity summary is composed of the evolution of properties and relations among these entities. To better understand our approach, we define the main concepts, i.e., RDF Molecule, Formal Concept Analysis and Fusion Policies, respectively.

Definition 3.1 (RDF Molecule [3]). If G is a given RDF Graph, we define an RDF Molecule M as a sub-graph of G such that,

$$M = \{t_1, \ldots, t_n\}, \forall (i,j) \in \{1, \ldots, n\}^2 \left(subject(t_i) = subject(t_j)\right)$$

where $t_1, \ldots, t_n$ denote the triples in M. In other words, an RDF Molecule M consists of triples which have the same subject. The RDF molecules are used as unit to produce the entity summary.

Definition 3.2 (Formal Concept Analysis [10]). aims at describing the relationships between objects and their attributes by considering binary data tables. In our approach, we apply the algorithm proposed by V. Vychodil [10], transforming RDF molecules into the binary data table it requires. Formal concepts are defined as conceptual clusters found within entity-property data tables. These data tables have rows corresponding to entities, and columns corresponding to the properties of those entities. Formal concepts are a set of $< A, B >$ pairs where A is the set of entities, B is a set of

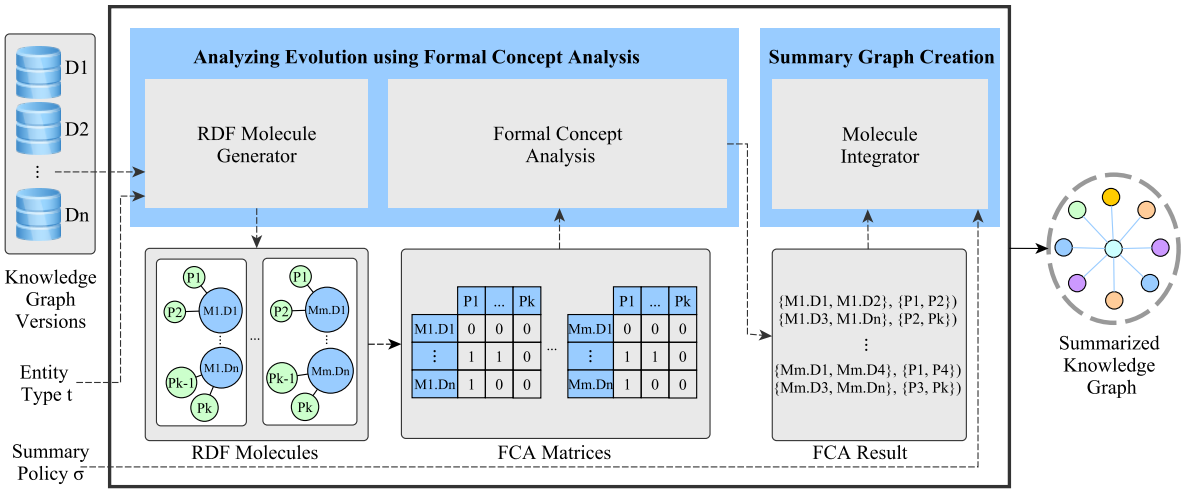


Figure 2: The approach receives a set of knowledge graph versions, an entity type, and a summary policy. The output is the entity summaries describing the evolution of knowledge among the different versions of the knowledge graph.

properties, and all the entities in A contain all the properties in B. A is known as *extent* and B is known as *intent*.

Definition 3.3 (Fusion Policy [1]). To produce a temporal evolution summary of entities spread over different versions of a knowledge graph, we resort to the concept of fusion policies defined by Collarana et al. [1]. A fusion policy is a set of rules operating on the triple level, which are triggered by a certain combination of predicates and objects. Fusion policies resort to an ontology O to resolve possible conflicts and inequalities on the levels of resources, predicates, objects and literals.

In this paper, we employ RDF molecules, Formal Concept Analysis and Fusion Policies to address the following problem: given a set of versions of RDF knowledge graphs, build an evolution summary of specific entities. The following section introduces and describes the architecture of our approach.

3.2 Architecture

Grounded on the semantic integration technique proposed by Collarana et al. [1], we propose a pipeline able to automatically create summaries of RDF entity evolution. Thus, a solution to *summarizing entity temporal evolution* out of different versions of an RDF knowledge graph is provided.

We propose a three-fold approach in order to identify equivalent entities in different versions of a knowledge graph and summarizing the evolution of those entities, thus providing a solution to the problem of *summarizing entity temporal evolution* in RDF knowledge graphs. This approach has three essential components. First, the pipeline takes input a set of knowledge graphs representing different temporal versions of the same knowledge graph. These graphs are then converted into a set of RDF molecules, from which molecules that represent the same real-world entity are grouped together. Each group of equivalent molecules are then converted into an $M \times N$ binary matrix. Second, the $M \times N$ matrix is supplied to the FCA component which performs formal concept analysis to summarize the temporal evolution of the molecules. Third, a summary fusion policy is applied to each output of the FCA component in order to produce a set of summary molecules. Each summary molecule represents the temporal evolution of a single entity over the knowledge graph versions taken as input. Figure 2 depicts the main components of this architecture.

3.3 Conversion of Knowledge Graphs to Groups of Equivalent RDF Molecules

The pipeline receives any number of Knowledge Graphs $\phi_1(D)$, $\phi_2(D)$,...,$\phi_n(D)$ as input where $1, 2, \ldots n$ represent the different temporal versions of the same Knowledge Graph $\phi(D)$. First each graph is individually converted into sets of RDF molecules. As defined in Preliminaries (see §3.1), RDF molecules consist of triples that have the same subject. Thus we obtain RDF molecule sets $S_1, S_2 \ldots S_n$ which correspond to graphs $\phi_1(D), \phi_2(D) \ldots \phi_n(D)$ respectively. The pipeline then identifies equivalent molecules within $S_1, S_2 \ldots S_n$. As $\phi_1(D), \phi_2(D) \ldots \phi_n(D)$ are different temporal versions of the same knowledge graph, it can be inferred that there exists equivalent molecules $M_1, M_2 \ldots M_n$ such that $M_1 \in S_1, M_2 \in S_2 \ldots M_n \in S_n$ and $M_1, M_2 \ldots M_n$ all represent the same real-world entity. For the sake of simplicity it is assumed here that equivalent entities retain the same URI. Practically, semantic similarity measures as demonstrated in [1] can also be integrated with this pipeline to identify equivalent entities in cases where the URI is different.

The equivalent RDF Molecules are then grouped together and each group is converted into a $M \times N$ matrix, where M corresponds to the number of molecules in the group and N corresponds to the number of distinct object or data property the molecules all contain. Each element in the matrix $e_{(i,j \to k)}$ represents whether molecule M_i contains a predicate j that has an object k. This value can be either `true` or `false` and is represented by a 1 and 0, respectively.

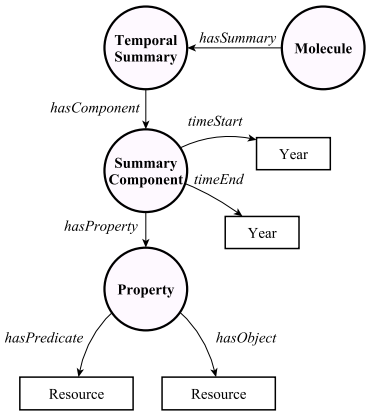


Figure 3: The ontology used to apply summary fusion policy for the creation of summary molecules.

Figure 1 presents a scenario where `Person.v2010`, `Person.v2012`, `Person.v2014` and `Person.v2016` all represent the same real world entity `Person` from Knowledge Graphs from year 2010, 2012, 2014 and 2016 respectively. In this first step of the pipeline, these molecules would be extracted from their respective knowledge graphs, identified as equivalent molecules, and then grouped together. This group of molecules would then be converted into a 4×7 matrix, 4 being the number of molecules in the group and 7 being the total number of distinct object/data properties in all the molecules. This matrix is shown in Table 1. The column headers, i.e. properties for this matrix are shown separately in Table 1. Once the $M \times N$ matrix has been created, it is then passed on to the next component in the pipeline to perform Formal Concept Analysis and summarize the temporal evolution in equivalent RDF entities.

3.4 Applying Formal Concept Analysis to Obtain a Summary of Evolution

As first mentioned in Definition 3.2, formal concept analysis studies binary object-attribute tables to describe the relationship between *objects* and their *attributes*. Our approach first converts knowledge graphs to RDF molecules. Within a single knowledge graph, an RDF molecule can be considered as an *object* while its object or data properties can be considered as *attributes*. When RDF molecules are modeled in this way, we are able to apply the formal concept analysis algorithm to compute formal concepts. The creation of these binary matrices is discussed in the previous section.

We apply the algorithm by V. Vychodil [10] on each group of RDF molecules obtained from the previous step. The algorithm returns a set of formal concepts $< M, P >$ where M is a set of all the molecules that have all the properties contained in P. In our approach the output $< M, P >$ from formal concept analysis gives us a set of molecules that have the same properties throughout different knowledge graph versions.

Table 1 demonstrates the input and output of formal concept analysis when applied to the RDF molecules in our motivation example. Table 1 shows the 4×7 binary matrix representing the

Algorithm 1: CreateSummary(FCAResult)

```
summary ← list() ;          // Initialize empty list
n ← FCAResult.length;
for j ←1 to n do
    (M, P) ← FCAResult[j];
    if M.length ≥ 1 and P.length ≥ 0 then
        summary ← summary + (M, P) ;
        FCAResult ← FCAResult − (M, P) ;
n ← FCAResult.length ;          // Recompute length
m ← summary.length;
for j ←1 to n do
    (M, P) ← FCAResult[j];
    if M.length1 and P.length ≥ 0 then
        for k ←1 to m do
            (X, Y) ← summary[k];
            if M[1] ∈ X then
                P ← P − Y
        if P.length ≥ 0 then
            summary ← summary + (M, P) ;
return summary
```

Figure 4: Algorithm for selecting distinct properties that form the *Summary Molecule*

object-attribute table. Table 1 shows the output of the FCA algorithm when supplied with this matrix. It supplies the information of which properties of the molecule has remained the same over which years. For example, Table 1 shows us that the entity Person had the same job (`hasJob` $\rightarrow$ *JobX*) and lived in the same city (`livesIn` $\rightarrow$ *CityA*) over the years 2010 and 2012.

The output of the FCA algorithm is supplied to the next step in order to create summary molecules.

3.5 Applying Fusion Policy to Integrate Summary into Knowledge Graph

To integrate the result obtained from the formal concept analysis algorithm, we resort to extend the fusion policies defined by Collarana et al. [1]. They defined the following: (1) the *Union Policy*, which includes all predicates-object pairs, removing the one that are syntactically the same; (2) the *Subproperty Policy*, which tracks if a property of one RDF molecule is an `rdfs:subPropertyOf` of a property of another RDF molecule, i.e., $\{r_1, p_1, A\}, \{r_2, p_1, B\} + O + subPropertyOf(p_1, p_2) \models \{\sigma_r(r_1, r_2), p_2, \sigma_v(A, B)\}$. As a result of applying this policy, the property p_1 is replaced with a more general property p_2; and (3) the *Authority Graph Policy*, which allows for defining one RDF graph as a prevalent source selecting its properties in case of property conflicts, i.e., properties annotated as `owl:FunctionalProperty`, equivalent properties `owl:equivalentProperty`, and equivalent classes annotated with `owl:equivalentClass` or `owl:sameAs`.

For the purpose of creating a *compact summary* of the temporal evolution of RDF molecules, we define an additional fusion policy

Table 1: Preparing equivalent RDF Molecules for Formal Concept Analysis: `Person` molecules from 4 different knowledge graph versions are grouped and compiled into a matrix for performing formal concept analysis. Rows correspond to each molecule, while columns correspond to distinct properties.

	P1	P2	P3	P4	P5	P6	P7
M1	1	0	1	0	1	0	0
M2	1	0	1	0	0	1	0
M3	0	1	0	1	0	1	0
M4	0	1	0	1	1	0	1

(a) FCA Input

Molecule	Property
M1	P1,P3,P5
M2	P1,P3,P6
M3	P2,P4,P6
M4	P2,P4,P5,P7
M1, M2	P1,P3
M2, M3	P6
M3, M4	P2,P4
M1, M4	P5

(b) FCA Output

	Molecule
M1	Person.v2010
M2	Person.v2012
M3	Person.v2014
M4	Person.v2016

(c) Molecules

	Property
P1	`livesIn` $\rightarrow$ *CityA*
P2	`livesIn` $\rightarrow$ *CityB*
P3	`hasJob` $\rightarrow$ *JobX*
P4	`hasJob` $\rightarrow$ *JobY*
P5	`isSpouseOf` $\rightarrow$ *P*
P6	`isSpouseOf` $\rightarrow$ *Q*
P7	`isFatherOf` $\rightarrow$ *M*

(d) Properties

called the *Summary Policy*. *Summary Policy* selects distinct properties from the output of the FCA algorithm, and follows the ontology shown in Figure 3 to create a *temporal summary graph*. A visual representation of the temporal summary graph according to our motivation example is shown in Figure 1e. For the sake of compactness, the summary graph is designed to have only one occurrence of each distinct property.

A *temporal summary graph* is created by first selecting all distinct properties out of the result obtained from the previous step. This is done by applying the algorithm shown in Figure 4 on a list of results like the one shown in Table 1. The algorithm takes as input the result R of the FCA algorithm. For every $(M, P) \in R, |M| = 1$, it returns (M, P') such that, $\forall (X, Y) \in R$ where $|X| > 1$ and $M \in X$, $P' = P \cap Y$ and $P' \neq \phi$.

For example, for the row [{M4}, {P2,P4,P5,P7}] in Table 1, the above algorithm is applied to obtain [{M4}, {P7}], i.e. [{Person.v2016}, {`hasChild` $\rightarrow M$}]. This can be seen in Figure 1e, where the property `hasChild` $\rightarrow M$ occurs only under the summary component of 2016. Every other property in Person.v2016 occurs also in other versions of the Person entity. This is represented in the summary graph under the summary components of (2010,2016) and (2014, 2016). The summary graph allows the possibility to reconstruct the original versions of the molecules. For example, in order to obtain the 2010 version of the Person molecule, it is only needed to select the edges of the summary graph that are annotated with the year 2010, and apply *Union Policy*.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we envision an approach for creating entity summaries automatically out of different temporal versions of a knowledge graph. To do so, the proposed approach utilizes the concepts of RDF molecules, Formal Concept Analysis, and Fusion Policies. We have explained the architecture and pipeline where only one parameter is needed, i.e., the entity type. The entity evolution summaries created by the approach may serve to create documentation, to display a visualization of the entity evolution, or to make an analysis of changes. As future work, we plan to implement and evaluate the approach considering its scalability and applicability.

ACKNOWLEDGEMENTS

This work is supported by the German Ministry of Education and Research (BMBF) in the context of the project MLwin ("Maschinelles Lernen mit Wissensgraphen", grant no. 01IS18050F).

REFERENCES

[1] Diego Collarana, Mikhail Galkin, Ignacio Traverso Ribón, Maria-Esther Vidal, Christoph Lange, and Sören Auer. 2017. MINTE: semantically integrating RDF graphs. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS*. 22:1–22:11. https://doi.org/10.1145/3102254.3102280

[2] Kemele M. Endris, Sidra Faisal, Fabrizio Orlandi, Sören Auer, and Simon Scerri. 2015. Interest-Based RDF Update Propagation. In *The Semantic Web - ISWC 2015*, Vol. 9366. 513–529. https://doi.org/10.1007/978-3-319-25007-6_30

[3] Javier D Fernández, Alejandro Llaves, and Oscar Corcho. 2014. Efficient RDF interchange (ERI) format for RDF data streams. In *International Semantic Web Conference*. Springer, 244–259.

[4] Tobias Käfer, Alexandra Wins, and Maribel Acosta. 2017. Modelling and Analysing Dynamic Linked Data using RDF and SPARQL. In *4th International Workshop on Dataset PROFIling and fEderated Search for Web Data (PROFILES) at the 16th International Semantic Web Conference (ISWC)*. CEUR-WS.

[5] Jan Michelfeit and Tomas Knap. 2012. Linked Data Fusion in ODCleanStore. In *ISWC Posters and Demonstrations Track*.

[6] Niko Popitsch and Bernhard Haslhofer. 2011. DSNotify – A solution for event detection and link maintenance in dynamic datasets. *Journal of Web Semantics* 9, 3 (sep 2011), 266–283. https://doi.org/10.1016/j.websem.2011.05.002

[7] Yannis Roussakis, Ioannis Chrysakis, Kostas Stefanidis, Giorgos Flouris, and Yannis Stavrakas. 2015. A flexible framework for understanding the dynamics of evolving RDF datasets. In *ISWC 2015 - International Semantic Web Conference*, Vol. 9366. 495–512. https://doi.org/10.1007/978-3-319-25007-6_29

[8] Anuj Singh, Rob Brennan, and Declan O'Sullivan. 2018. DELTA-LD: A Change Detection Approach for Linked Datasets. In *4th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW)*. http://ceur-ws.org/Vol-2112/mepdaw{_}paper{_}1.pdf

[9] Jürgen Umbrich, Boris Villazón-Terrazas, and Michael Hausenblas. 2010. Dataset Dynamics Compendium: A Comparative Study. In *Proceedings of the First International Workshop on Consuming Linked Data (COLD2010)*. http://oa.upm.es/6632/

[10] Vilem Vychodil. 2008. *A new algorithm for computing formal concepts.* na.

Open traffic lights: a strategy for publishing and preserving traffic lights data

Brecht Van de Vyvere
IDLab
Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium,
brecht.vandevyvere@ugent.be

Pieter Colpaert
IDLab
Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium,
pieter.colpaert@ugent.be

Erik Mannens
IDLab
Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium,
erik.mannens@ugent.be

Ruben Verborgh
IDLab
Department of Electronics and Information Systems
Ghent University – imec, Ghent, Belgium
ruben.verborgh@ugent.be

ABSTRACT

For better traffic flow and making better policy decisions, the city of Antwerp is connecting traffic lights to the Internet. The live "time to green" only tells a part of the story: also the historical values need to be preserved and need to be made accessible to everyone. We propose (i) an ontology for describing the topology of an intersection and the signal timing of traffic lights, (ii) a specification to publish these historical and live data with Linked Data Fragments and (iii) a method to preserve the published data in the long-term. We showcase the applicability of our specification with the opentrafficlights.org project where an end-user can see the live count-down as well as a line chart showing the historic "time to green" of a traffic light. We found that publishing traffic lights data as time sorted Linked Data Fragments allow synchronizing and reusing an archive to retrieve historical observations. Long-term preservation with tape storage becomes feasible when archives shift from byte preservation to knowledge preservation by combining Linked Data Fragments.

This is a print-version of a paper first written for the Web. The Web-version is available at https://brechtvdv.github.io/Article-Open-Traffic-Lights

CCS CONCEPTS

• Information systems~Web interfaces • Applied computing~Digital libraries and archives

KEYWORDS

smart cities, intelligent agents, linked data, semantic web

ACM Reference format:

Brecht Van de Vyvere, Pieter Colpaert, Erik Mannens and Ruben Verborgh. 2019. Open traffic lights: a strategy for publishing and preserving traffic lights data. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316520

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316520

1 Introduction

Smart cities make evidence-based decisions based on, among others, sensors installed in the city. As government data and policy decisions world-wide, such as open by default, the freedom of information act or the public sector information directive, this data must be made freely available to everyone. In the case of sensor data, not only the live values are important, also historical values need to be preserved and made available. When historic values of smart city datasets become available, everyone can join an evidence-based debate on the city's future.

The first traffic light was installed in London in 1868. Today, Traffic Control Systems (TCS) do not differ much from the model installed 1.5 centuries ago, yet we can imagine the potential of making TCSs internet-connected and making traffic light statuses available as Open Data. Not only self-driving cars may benefit, also cyclists seeking green lights on a rainy day, or a citizen willing to analyze the crossroad's efficiency near their home. In the Netherlands, hundreds of these TCSs are already converted through the public-private coorporation Talking Traffic. A centralized approach using asynchronous messaging is implemented with the Traffic Live Exchange (TLEX) system which distributes data between TCSs and the end-user in both directions. The data so far has not been published publicly, but only been made available in partner applications for end-users. In the city of Antwerp (Belgium), one geographical zone is

accommodated with different kinds of sensors to conduct experiments, such as the publication of its traffic lights data that is described in this paper.

Traffic lights data standards are already created by the European Telecommunications Standards Institute (ETSI), yet identifiers are scoped to the intersection, region or country. With Linked Data (Section 2), traffic light streams may become interoperable worldwide and may allow for a decentralized data publishing strategy. This way, any client, whether it is a route planners, a self-driving car or a data analyst, can follow links to and from other datasets while querying.

Preserved data from among others these TCSs can be used by digital scholars, citizens, journalists, and city officials… to study the city's digital pulse. Such Web archives are already used by scholars in, among others, the social sciences and humanities for studying phenomena [1]. In this paper, we propose a publishing strategy for data from TCSs, and zoom in on the aspect of preservation.

The state of the art is described in Section 2. In Section 3, we propose a method how these archives also can preserve sensor observations and allow linking from the Open Data publisher to the archive. Next, an end-user application with code-snippet is demonstrated in Section 4. Finally, we discuss our conclusions in Section 5.

2 Background

2.1 Preservation

Digital content can be preserved on long-term on tape or on disk storage. In general, tape is used by TV broadcasters or audiovisual archives, because of its low pricing for big volumes (>50 terabytes). A tape drive is cost-effective when writing or reading large files, because the position of the tape drive reader does not need to be moved frequently contrary to small files. This deficit is solved by either bundling small files together (e.g. zip, tar) or saving them on disk storage. The latter benefits over tape with faster search and retrievable times and the ability to preserve small files more efficiently.

To archive a document, this is traditionally done by generating a hash (e.g. MD5) of the document, copying the document together with a metadata file containing the hash (cfr. sidecar) to the archive system and verify the content by regenerating the hash on the archive.

2.2 Resource Description Framework

The Resource Description Framework (RDF) is a graph-based model to describe statements. Every RDF statement corresponds with a triple (subject-predicate-object): the *subject* is a resource that something is said about. The subject is related to an *object* through a *predicate*. For example, 'the traffic light is lighted green': 'the traffic light' is the subject, 'green' is the object and 'is lighted' is the predicate that expresses the relation.

2.3 Memento TimeMaps

Memento [2] is a HTTP framework that allows a client to retrieve a time specific version (called a Memento) of an Original Web resource that exists or used to exist. One of the techniques to do this is by retrieving a TimeMap that describes a list of Mementos of the Original Web resource with its valid timestamps.

2.4 Linked Data Fragments

Linked Data Fragments [3] is a conceptual framework that allows to compare Linked Data Web interfaces (a datadump, SPARQL-endpoint, subject page…) by describing how they publish fragments of their RDF datasets. Such a Linked Data Fragment consists of three parts:

- a **selector** defines which parts of the datasets belong to a fragment. For example: only sensor observations that are generated between a certain time interval.
- **metadata** about the fragment. For example: the license that is applied.
- **controls** help clients to retrieve more relevant data. For example: where a previous page of a paged collection can be found.

The Linked Data Fragments axis shows a uniform view over the trade-offs (caching, bandwidth etc.) that these Web interfaces bring in respect to the client and server effort. For instance, an interface that offers a datadump requires a higher effort from the client than the server, because the client needs to download the whole dataset before it can solve its task.

Comunica [4] is a framework to query over the various Web interfaces in the Linked Data Fragments axis. It interprets the metadata and controls to evaluate a query. Finding a fragmentation strategy [5] for Web archives and time series data that can be consumed by generic clients [6] is still a challenge. In this paper we try to suggest a solution in which generic clients can still query over the data.

3 Open Traffic Lights

The Open Traffic Lights project started from the need to publish the traffic lights data in the Smart Zone in Antwerp as Open Data. By using the Resource Description Framework (RDF) as a common model for representing knowledge, sensor observations are syntactically interoperable without necessarily having to support the same serialization. To enable this for traffic lights observations, an ontology is needed to describe the signal phase and timing and how traffic can move over the intersection.

The map data describing the corss road and the signal phase and timing of the traffic control messages are available through a closed message broker (cfr. MQTT) with a frequency of 200 milliseconds which corresponds with the cycle time of the Traffic Control System of the intersection. In the specification section, we will describe how the number of data messages can be optimized for Open Data publishing and

how historical information is still retrievable, in contrast to the broker interface.

The ontology is based on two Flemish profiles: the MapData (MAP) and the Signal Phase and Timing (SPAT) profile. While MAP focuses on describing the topology of an intersection, SPAT describes the timing of the phase that a traffic light adheres to. Both profiles conform to European Telecommunications Standards Institute (ETSI) standards: ETSI 103 301, ETSI TS102 894-2, but are not publicly available.

3.1 Ontology

The Open Traffic Lights ontology is currently scoped to the data that is published by the Smart Zone. It is publicly available under an Open license: https://w3id.org/opentrafficlights# with prefix otl. Fig. 1 shows an overview of all the supported classes and properties that are defined.

driv

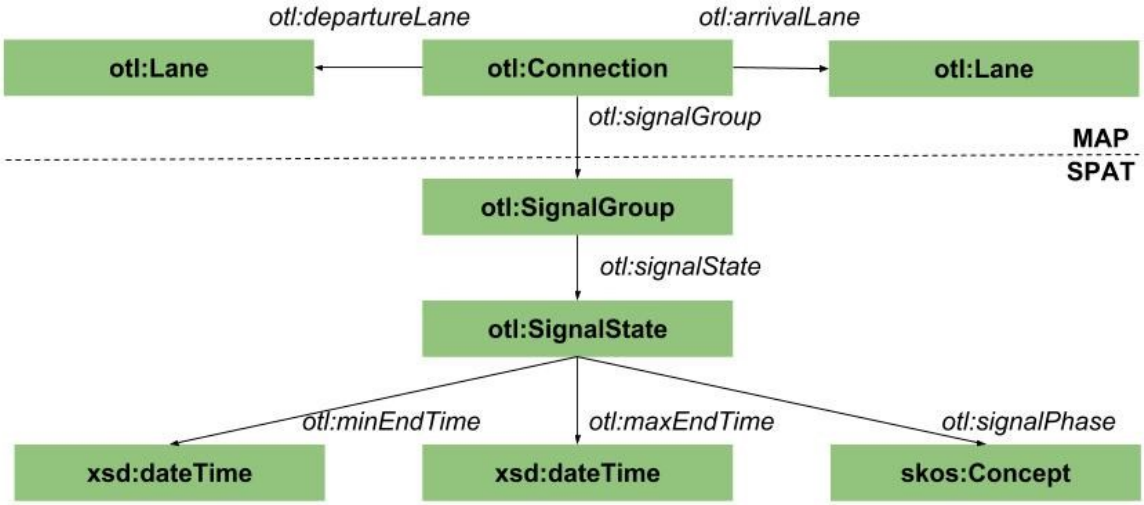


Fig. 1: Classes and properties for describing MAP/SPAT related knowledge.

According to the driving direction of the road user, the lane *otl:Lane* that goes towards the conflict area of the intersection is described with a *otl:departureLane*, mutatis mutandis for the *otl:arrivalLane*. If a road user can travel from a departure lane towards an arrival lane, than this is called a *otl:Connection*.

Travel is possible according to the *otl:SignalState* of the traffic light. Instead of referring to a traffic light, the SPAT profile coins the term *otl:SignalGroup* for the collection of traffic lights that continuously share a signal state. The latter is characterized by the following relations:

- **otl:signalPhase**: points to a SKOS concept that represents the signal phase (stop and remain, movement allowed, unavailable…). As a result of this phase, the traffic light shows a color (green, orange flashing, etc.).
- **otl:minEndTime**: the earliest time that the signal phase will change.
- **otl:maxEndTime**: the maximum time that the signal group will remain in this signal phase.

In the SPAT profile, the signal phase is described using a text field, e.g. "protected-Movement-Allowed", which means that a road user can safely cross the intersection (a green light). To allow semantic interoperability with foreign countries, we made a taxonomy of possible signal phases which can be dereferenced by HTTP clients. The taxonomy is publicly available at https://w3id.org/opentrafficlights/thesauri/signalphase under an Open license.

In next section, we will describe a Web Application Programming Interface specification to publish traffic light observations for Open Data re-use.

3.2 Specification

The specification contains three aspects:

- how every traffic light observation must be described,
- how the server interface exposes the live observations,
- how historical observations must be published

A data publisher *must* publish observations with instantaneous graphs (iGraphs) [7] which represent a part of the traffic lights stream. This iGraph is a named graph that annotates one or more observations with a time stamp. A publisher *must* use *prov:generatedAtTime* to indicate the iGraph's timestamp. Listing 1 shows an iGraph https://opentrafficlights.org/spat/K648?time=2018-10-31T14:58:23.205Z that contains the signal phase and timing of signal group https://opentrafficlights.org/id/signalgroup/K648/6. The iGraph *should* be extended with the Sensor, Observation, Sample, and Actuator [8] (SOSA) ontology to semantically describe an observation with its result, observed property etc.

Listing 1: Example of an iGraph *https://opentrafficlights.org/spat/K648?time=2018-10-31T14:58:23.205Z*which time annotates the state of a signal group. The state *_:b0* links to a phase (*https://w3id.org/opentrafficlights/thesauri/signalphase/0*) and is constrained with a time interval when it will end.

```
<https://opentrafficlights.org/id/signalgroup/K648/6>a otl:SignalGroup.
<https://opentrafficlights.org/spat/K648?time=2018-10-31T14:58:23.205Z>
{
    <https://opentrafficlights.org/id/signalgroup/K648/6>
otl:signalState _:b0.
    _:b0 a otl:SignalState;
    otl:signalPhase
<https://w3id.org/opentrafficlights/thesauri/signalphase/0>;
    otl:minEndTime "2018-10-31T14:58:37.605Z"^^xsd:dateTime;
    otl:maxEndTime "2018-10-31T14:59:40.605Z"^^xsd:dateTime.
}
<https://opentrafficlights.org/spat/K648?time=2018-10-31T14:58:23.205Z>
prov:generatedAtTime "2019-02-14T17:43:54.297Z"^^xsd:dateTime.
```

SPAT messages are sent every cycle time of a Traffic Control System. For the TCS in Antwerp, which has a cycle time every 200 ms, this corresponds with 5 iGraphs per second. For our use case whole second resolution suffices as we want to show a live count-down based on seconds. To lower the number of messages, a server *must* calculate the final outcome that an end-user will see and only publish if this changes. A new observation therefore is published whenever the minimum and maximum count-down (now - otl:minEndTime/otl:maxEndTime) in seconds or the otl:signalPhase changes. This results in the generation of one iGraph per second.

There are two main strategies to publish live data: a publish/subscribe system where an iGraph is pushed to the client and HTTP polling where the client pulls a Linked Data Fragment repeatedly. The data publisher *must*offer HTTP polling and *should* offer pub/sub:

- **HTTP polling**: a Linked Data Fragment which is a HTTP document containing one or more of the most recent iGraphs (e.g. https://lodi.ilabt.imec.be/observer/rawdata/latest). The *ETag* header *must* be added to indicate if the content has changed so HTTP caching is possible. This ETag can be generated by hashing (e.g. MD5) the fragment or the latest iGraph.
- **publish/subscribe**: every update corresponds with one iGraph like Listing 1. Preferably, HTTP-based Server-Sent Events *should* be chosen.

Historical data are published as Linked Data Fragments similar to the HTTP polling approach. Fig. 3 shows that all observations are ordered on a time axis: the Linked Data Fragments **select** data based on time ranges. The historic fragments *should* have a *cache-control* header set to *immutable* instead of using an ETag header. To allow a HTTP client to automatically discover older observations, a *hydra:previous* **hypermedia** link *must* be added using the Hydra vocabulary. *Hydra:next* links *should* be added to allow time range retrieval in both directions. Also, the URL of the fragment is identified using the datetime of its first observation (e.g. https://opentrafficlights.org/spat/K648?time=2018-10-31T14:58:23.205Z).

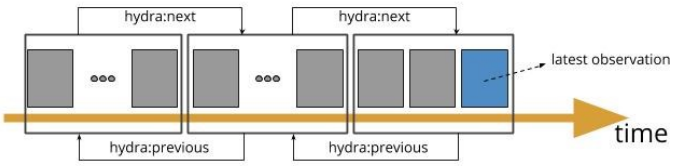


Fig. 2: Timeseries are published as a paged collection of time sorted Linked Data Fragments.

Another feature that the interface should expose is templated links: a client should be able to construct a URL to retrieve the fragment that contains observations around a certain time parameter. Fig. 4 gives an example on how to express this hypermedia control. The server must redirect with a HTTP 302 status code to the fragment whose time range encapsulates time. When missing, the last fragment is returned. With the templated links hypermedia control, Open Data reusers don't have to crawl through all the fragments to fetch observations between a certain time range. Technically, templated links can replace pagination hypermedia links, but these links are necessary to lower the barrier for Open Data reusers such as regular Web crawlers [9].

Listing 2: A client can search for observations with a datetime as input parameter.

```
{
      "@id": "«currentPageURL»",
      "@type": "hydra:PartialCollectionView",
      "hydra:search": {
            "@type": "hydra:IriTemplate",
            "hydra:template": "<<baseUrl>>{?time}",
            "hydra:variableRepresentation": "hydra:BasicRepresentation",
            "hydra:mapping": {
                  "@type": "hydra:IriTemplateMapping",
                  "hydra:variable": "time",
                  "hydra:required": true
            }
      }
}
```

Finally, following **metadata** *must* be added to every fragment:

- the **topology** of the intersection (MAP): when available, the identifiers of the lanes of the local authorities should be reused. Otherwise, a lane should be defined using *otl:Lane* and use http://purl.org/dc/terms/description and http://www.opengis.net/#geosparql/wktLiteral so user agents can discover which travels are possible and on which signal group this depends (*otl:signalGroup*).
- an Open **License** (cfr. CC-0, CC-BY etc.) is needed so the data complies under the Open definition and that reusers can check whether its intentions with the data are in line with the legal terms

In next section, a method will be proposed for long-term preservation of these fragments.

3.3 Preservation strategy

With our specification, one observation with the average size of 7 kilobytes is published every second which generates more than a half gigabyte per day. It is important that inactive data is removed from the Open Data server to keep it light-weight, however, these should still be available for retrieval.

As discussed in the background, tape is cost-efficient for large files. To preserve the published fragments, which are relatively small, on tape a different approach than disk storage is required. To circumvent this, Linked Data Fragments can be merged into one bigger Linked Data Fragment. This means that not the exact bytes are preserved, but only the RDF statements. On the one hand, when there are no blank nodes used in the statements, it is easy to check whether the merged fragment contains all the statements of the original fragments. On the other hand, when there are blank nodes, the merged fragment needs to rewrite the original blank nodes so these remain locally unique. For disk archives, there is no need to create bigger files, thus archiving can be done traditionally as explained in section 2.

There are four aspects on synchronizing archives with our specification for traffic lights data (Fig. 3):

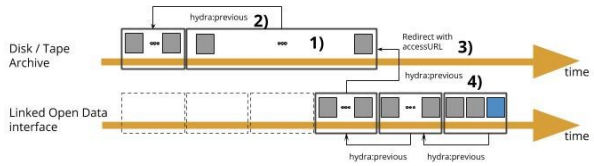


Fig. 3: An archive can harvest the historical observations, optionally merge the Linked Data statements of multiple documents into one document, and link everything together through previous links.

- archives need to harvest and optionally merge the Linked Data Fragments by following the provided hypermedia links in the Linked Open Data interface
- a hypermedia link is added to its previous harvested fragment
- an access URL is exposed by the archive that redirects to the latest harvested fragment
- the last published fragment of the Linked Open Data interface links to this access URL so the archive becomes an extension

4 Demonstrator

This application shows live traffic lights data of the intersection in Antwerp. The source code can be found as a Codepen at https://codepen.io/brechtvdv/pen/BMQPNX/. The dataset is freely available under an Open License at https://lodi.ilabt.imec.be/observer/rawdata/latest.

First, a user can select a departure and arrival lane, indicating how the intersection will be crossed. The departure and arrival lane describe an *otl:Connection* that can be selected from this information. A live count-down of the traffic light that is responsible for the chosen otl:Connection is shown. The count-down is calculated with the minimum ending time of the state (current time - otl:minEndTime). The bigger then '>' sign indicates that the minimum and maximum end time are not equivalent, thus it is uncertain when the signal state ends. Note that this application uses HTTP polling every 100 milliseconds.

Underneath the count-down, a historic 'time-to-green' is shown. The line chart represents how many seconds a user had to wait before getting a green light. The horizontal lines correspond with a waiting time of 0 seconds, and thus corresponds with a green light.

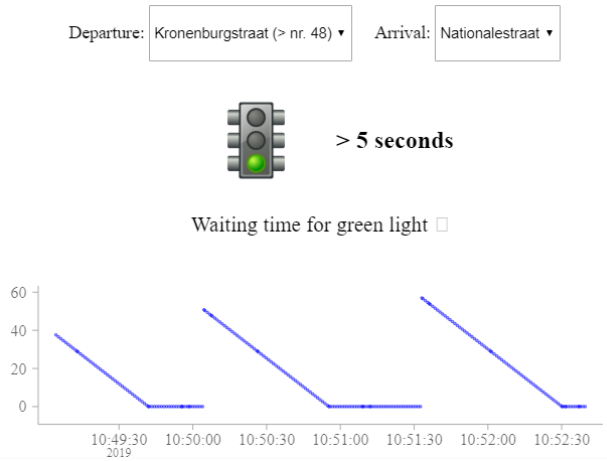


Fig. 4: Visualisation of the live count-down and the amount of seconds to wait before having a green light. Notice that the horizontal lines correspond with green times.

5 Conclusion

Cities can use the Open Traffic Lights ontology to publish their traffic lights data in a semantically interoperable way. A TCS Open Data interface following the proposed specification publishes every second one update. This lowers the barrier for Open Data re-users to create visualizations on top of it: they only have to retrieve the latest observation as fast as possible. Still, it is possible to create a more intelligent client that calculates a client-side count-down. While this can improve the user experience, another information resource would have to be created to inform about priority updates (such as when an ambulance or police car influences the plan of the TCS).

This early work is open to feedback from other cities that want to publish TCS data. In earlier work, we proposed a Linked Times Series interface [5] as part of the publishing strategy for Flanders. In this paper, archives are able to download the historic data of cities through its hypermedia links, and optimize storage for tape storage. By linking the archived fragments in a similar way the archive is activated as an Open Data extension of the interface. In future work, we will work on improving the discoverability of traffic lights datastreams by applying the Semantic Sensor Network [10] and the Vocabulary and Catalog for Linked Streams [11] ontologies. Also, we will explore how Memento TimeMaps compare to our time sorted Linked Data Fragments solution. Furthermore, we will benchmark HTTP polling versus publish/subscribe to have a better understanding how the speed of sensor observations impacts the server

infrastructure needed to guarantee and acceptable end-user latency for the live updates.

6 ACKNOWLEDGMENTS

The described research activities were funded by Ghent University, imec, Flanders Innovation & Entrepreneurship (AIO) and supports the Observer (imec City of Things) project.

REFERENCES

[1] Vlassenroot, et al., Eveline: Web archives as a data resource for digital scholars. In: International Journal of Digital Humanities. Springer (2019).
[2] Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: Time travel for the web. arXiv preprint arXiv:0911.1112. (2009).
[3] Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., Colpaert, P.: Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web. Journal of Web Semantics. 37–38, (2016).
[4] Taelman, R., Van Herwegen, J., Vander Sande, M., Verborgh, R.: Comunica: a Modular SPARQL Query Engine for the Web. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E. (eds.) Proceedings of the 17th International Semantic Web Conference. pp. 239–255. Springer (2018).
[5] Rojas Meléndez, J.A., Van de Vyvere, B., Gevaert, A., Taelman, R., Colpaert, P., Verborgh, R.: A Preliminary Open Data Publishing Strategy for Live Data in Flanders. In: WWW2018, the International World Wide Web Conference. pp. 1847–1853. ACM Press (2018).
[6] Taelman, R., Verborgh, R., Colpaert, P., Mannens, E.: Moving Real-Time Linked Data Query Evaluation to the Client. In: European Semantic Web Conference. pp. 3–7. Springer (2016).
[7] Barbieri, D.F., Valle, E.D.: A proposal for publishing data streams as linked data. In: Linked Data on the Web Workshop (2010).
[8] Janowicz, K., Haller, A., Cox, S.J.D., Le Phuoc, D., Lefrançois, M.: SOSA: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics. (2018).
[9] Niu, J.: An overview of web archiving. D-Lib magazine. 18, (2012).
[10] Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro RaúL, Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., others: The SSN ontology of the W3C semantic sensor network incubator group. Web semantics: science, services and agents on the World Wide Web. 17, 25–32 (2012).
[11] Tommasini, R., Sedira, Y.A., Dell'Aglio, D., Balduini, M., Ali, M.I., Le Phuoc, D., Della Valle, E., Calbimonte, J.-P.: VoCaLS: Vocabulary and Catalog of Linked Streams. In: International Semantic Web Conference. pp. 256–272. Springer (2018).
[12] Vlassenroot, et al., Eveline: Web archives as a data resource for digital scholars. In: International Journal of Digital Humanities. Springer (2019).
[13] Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: Time travel for the web. arXiv preprint arXiv:0911.1112. (2009).
[14] Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., Colpaert, P.: Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web. Journal of Web Semantics. 37–38, (2016).
[15] Taelman, R., Van Herwegen, J., Vander Sande, M., Verborgh, R.: Comunica: a Modular SPARQL Query Engine for the Web. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E. (eds.) Proceedings of the 17th International Semantic Web Conference. pp. 239–255. Springer (2018).
[16] Rojas Meléndez, J.A., Van de Vyvere, B., Gevaert, A., Taelman, R., Colpaert, P., Verborgh, R.: A Preliminary Open Data Publishing Strategy for Live Data in Flanders. In: WWW2018, the International World Wide Web Conference. pp. 1847–1853. ACM Press (2018).
[17] Taelman, R., Verborgh, R., Colpaert, P., Mannens, E.: Moving Real-Time Linked Data Query Evaluation to the Client. In: European Semantic Web Conference. pp. 3–7. Springer (2016).
[18] Barbieri, D.F., Valle, E.D.: A proposal for publishing data streams as linked data. In: Linked Data on the Web Workshop (2010).
[19] Janowicz, K., Haller, A., Cox, S.J.D., Le Phuoc, D., Lefrançois, M.: SOSA: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics. (2018).
[20] Niu, J.: An overview of web archiving. D-Lib magazine. 18, (2012).
[21] Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro RaúL, Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., others: The SSN ontology of the W3C semantic sensor network incubator group. Web semantics: science, services and agents on the World Wide Web. 17, 25–32 (2012).
[22] Tommasini, R., Sedira, Y.A., Dell'Aglio, D., Balduini, M., Ali, M.I., Le Phuoc, D., Della Valle, E., Calbimonte, J.-P.: VoCaLS: Vocabulary and Catalog of Linked Streams. In: International Semantic Web Conference. pp. 256–272. Springer (2018).

MisInfo'19 Chairs' Welcome

It is our great pleasure to welcome you to the *First International Workshop on Misinformation, Computational Fact-Checking and Credible Web* at TheWebConf 2019. It is a recognized problem that our society is struggling with an unprecedented amount of falsehood that harms wealth, democracy, and health. Debunking misinformation and disinformation calls for interdisciplinary collaboration of and advancements in multiple areas, including journalism, communication studies, law and public policy, psychology, and political science. Computing technology plays a crucial role in it. The last few years have witnessed a substantial growth in efforts at computational fact-checking, of which many are data-driven, AI-powered, and include human in the loop. These efforts tackle various fronts, such as the detection of fabricated news, rumors, and spam on social media, automation in fact-checking, flagging clickbait articles, and discovering fake accounts and malicious social media bots, but are far from the maturity that is required to reach commercial application.

The objective of this workshop is to bring together the researchers, practitioners, and educators that take computational and data-driven approaches to address the above challenges.

The workshop received 19 submissions, out of which 9 were accepted. These papers cover a broad range of topics: how to manage crowds and professionals in the fact-checking process, analysis of partisan behavior of social bots, automatic identification of worth-checking sentences and of unreliable news sources, a data model linking facts with beliefs, a study of misinformation in WhatsApp groups, how information security and counter-disinformation strategies apply to misinformation.

We succeeded in attracting prominent invited speakers on the credibility of information online and on the psychological and interpersonal processes in social media. Thanks to the support from Google and Facebook, we are also been able to offer free registrations to students.

We believe that the program of Misinfo 2019 provides a good balance between several topical areas and that it will provide attendees a valuable opportunity to learn and share ideas with other researchers.

Laks V.S. Lakshmanan
University of British Columbia, Canada

Chengkai Li
University of Texas at Arlington, USA

Paolo Papotti
EURECOM, France

MisInfo'19 Organization

Workshop Chairs:
Laks V.S. Lakshmanan (University of British Columbia, Canada)
Chengkai Li (University of Texas at Arlington, USA)
Paolo Papotti (EURECOM, France)

Steering Committee:
Bill Adair (Sanford School of Public Policy, Duke University, USA)
Juliana Freire (New York University, USA)
Laks V.S. Lakshmanan (University of British Columbia, Canada)
Chengkai Li (University of Texas at Arlington, USA)
Ioana Manolescu (Inria Saclay, France)
Alexios Mantzarlis (Poynter Institute, USA)
Preslav Nakov (Qatar Computing Research Institute, Qatar)
Paolo Papotti (Eurecom, France)
Jun Yang (Duke University, USA)
Cong Yu (Google Research, USA)

Program Committee:
Jisun An (Qatar Computing Research Institute, Qatar)
Mevan Babakar (Full Fact, UK)
Alberto Barron-Cedeno (Qatar Computing Research Institute, Qatar)
Simon Baumgartner (Google, USA)
James Caverlee (Texas A&M University, USA)
Giovanni Da San Martino (Qatar Computing Research Institute, Qatar)
Nicholas Diakopoulos (Northwestern University, USA)
Peter Fray (University of Technology Sydney, Australia)
Luis Gravano (Columbia University, USA)
James T. Hamilton (Stanford University, USA)
Naeemul Hassan (University of Mississippi, USA)
Srijan Kumar (Stanford University, USA)
Angela Lee (University of Texas at Dallas, USA)
Shirin Nilizadeh (University of Texas at Arlington, USA)
Deokgun Park (University of Texas at Arlington, USA)
Craig Silverman (Buzzfeed, USA)
Jonathan Stray (School of Journalism, Columbia University, USA)
Xavier Tannier (Sorbonne University, France)
Mark Tremayne (University of Texas at Arlington, USA)
Immanuel Trummer (Cornell University, USA)
Will Wu (Google, USA)
Gensheng Zhang (Google, USA)
Arkaitz Zubiaga (University of Warwick, UK)

Looking Backward to See a Way Forward

How Credibility Research in Social Science Can Help in the Battle Against Misinformation

Miriam J. Metzger
Department of Communication
University of California Santa Barbara, Santa Barbrara, CA, USA, metzger@ucsb.edu

ABSTRACT

Credibility research in the social sciences has a long history that may be particularly informative to today's efforts to combat misinformation and disinformation online. This keynote address will discuss the ways in which credibility has been studied in the disciplines of communication and psychology, including both how this notion has been conceptually and operationally defined. Key research findings will be presented with the aim of understanding what kinds of computational algorithms, tools, systems, and applications for tackling misinformation and disinformation are more versus less likely to be effective. This will help to answer a very important yet often overlooked question, which is: To what extent is misinformation a problem of information or a problem of human information processing? The answer to this question is crucial for both software developers and designers of educational intervention efforts to minimize the negative societal impacts of fabricated news and information flowing over the Internet and on social media.

CCS CONCEPTS

• Applied computing~Psychology

KEYWORDS

Credibility; social science; misinformation; fake news; social media

ACM Reference format:

Miriam J. Metzger. 2019. Looking Backward to See a Way Forward: How Credibility Research in Social Science Can Help in the Battle Against Misinformation. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317602

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317602

A Topic-Agnostic Approach for Identifying Fake News Pages

Sonia Castelo
New York University
s.castelo@nyu.edu

Thais Almeida
Federal University of Amazonas
tga@icomp.ufam.edu.br

Anas Elghafari
New York University
anas.elghafari@nyu.edu

Aécio Santos
New York University
aecio.santos@nyu.edu

Kien Pham
New York University
kien.pham@nyu.edu

Eduardo Nakamura
Federal University of Amazonas
nakamura@icomp.ufam.edu.br

Juliana Freire
New York University
juliana.freire@nyu.edu

ABSTRACT

Fake news and misinformation have been increasingly used to manipulate popular opinion and influence political processes. To better understand fake news, how they are propagated, and how to counter their effect, it is necessary to first identify them. Recently, approaches have been proposed to automatically classify articles as fake based on their content. An important challenge for these approaches comes from the dynamic nature of news: as new political events are covered, topics and discourse constantly change and thus, a classifier trained using content from articles published at a given time is likely to become ineffective in the future. To address this challenge, we propose a topic-agnostic (TAG) classification strategy that uses linguistic and web-markup features to identify fake news pages. We report experimental results using multiple data sets which show that our approach attains high accuracy in the identification of fake news, even as topics evolve over time.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**; Feature selection.

KEYWORDS

Misinformation; Fake News Detection; Classification; Online News

ACM Reference Format:
Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316739

1 INTRODUCTION

Fake news have been increasingly used to manipulate public opinion and influence political processes. This has been made possible both by the existing Web infrastructure and online media platforms (e.g., Facebook, Twitter), which make it easy for information to be propagated. Unlike traditional print media, information can be published on Web sites and shared among users in social media platforms with no third party filtering or fact-checking [1]. Given that 62% of adults in the US consume news from social media [8] and many who see fake stories report that they believe them [21], these platforms have become a target for propaganda campaigns [2, 11].

While fake news have attracted substantial attention, the problem is not well understood [11]. It is challenging to discover and cross-reference conflicting sources and claims. This problem is compounded due to the large number of news sites and high volume of content. Automated methods that identify potential fake news and unreliable news sources can aid manual fact checking by providing contextual information and limiting the volume of content that the human fact-checker needs to consider. Such methods can also help us better understand the ecosystem of fake news: where they start, how they propagate, and how to counter their effects.

However, the automatic identification of fake news is a hard problem, given that news cover a wide variety of topics and linguistic styles, and can be shared on many different platforms [18].

In this paper, we study the problem of detecting fake news published on the Web. Given a web page P, our goal is to determine whether P is likely to contain fake news. Since some sites publish a mix of fake and real news [1], we consider pages published by suspicious sites as unreliable, and pages published by legitimate media outlets as reliable. We note that there is no widely-accepted definition for fake news. Here, we focus on all types of active political misinformation that go beyond old-fashioned partisan bias, and consider unreliable not only sites that publish fabricated stories, but also sites that have a pattern of misleading headlines, thinly-sourced claims, and that promote conspiracy theories. We exclude from our definition satire sites as well as opinion sites – even if extreme – if they do not display a pattern of promoting misinformation.

Previous approaches to fake news identification have largely focused on using the content of news web pages to determine their veracity [14, 17]. However, using the content has important limitations. Notably, given the dynamic nature of news, as new events are covered, topics and discourse constantly change and thus, a classifier trained using content from articles published at a given time is likely to become ineffective in the future. In addition,

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316739

studies have found that page content alone is not sufficient to accurately classify the veracity of news [4, 6, 22].

Our Approach. To address this limitation, we *propose a new classification strategy that is topic-agnostic*. Instead of using the bag of words in a page, we explore topic-agnostic features, including web-markup and linguistic features that are common in fake news.

We perform a *detailed experimental evaluation* using publicly-available datasets [1, 14] and a new dataset we created – the PoliticalNews dataset. Since existing datasets contain a small number of articles or cover a narrow time span, they are insufficient to assess the topic-agnosticism aspect of our approach. The PoliticalNews dataset contains a mix of political topics and spans several years. We report results which show that our approach is effective, obtaining accuracies that are between 8% and 24% higher than the baseline, and it is robust as topics change over time as well when applied to different domains.

Contributions. Our main contributions are: (i) we propose a classification strategy for identifying fake news which, to the best of our knowledge, is the first to rely solely on topic-agnostic features; and (ii) we present the results of an experimental evaluation, using multiple datasets and considering various baselines, which show that our approach is robust and attains high accuracy.

2 RELATED WORK

In this section, we discuss techniques that have been proposed to detect fake news published on Web sites. We also discuss publicly-available datasets that have been used to evaluate these techniques.

Detecting Fake News on the Web. Potthast et al. [16] investigated the use of writing style which included features such as n-grams and readability scores. They found that while style-based and topic-based classifiers are effective at differentiating hyper-partisan news from mainstream news (0.75 accuracy), they are not effective at differentiating fake from real news (0.55 accuracy). Horne and Adali [9] used linguistic features including readability scores, sentence structure, and part of speech of the words used. These features were very effective for the task of differentiating satire from real news (0.91 accuracy), but somewhat less so for differentiating fake news from real (0.78 accuracy). Pérez-Rosas et al. [14] also considered writing style and proposed the use of features that capture content-based aspects of web pages, such as n-grams, punctuation, psycho-linguistic, readability and syntax. Their model attains accuracy up to 0.76. Fairbanks et al. [6] investigated whether credibility and bias can be assessed using content-based and structure-based methods. The structure-based method constructs a reputation graph where each node represents a site, and the edges represent mutually linked sites, as well as shared files. This work shares some elements with our work in its usage of features derived from the HTML source of the pages (they use a subset of the feature in our classifier), but their focus is on the network between sites.

The picture that emerges from these approaches is that content-based features, while effective for detection of bias and satire, often fall short for detecting fake news. Some of works that have achieved good results on fake news detection have done so by including additional information about the sites. In our approach, we follow a similar direction and examine the combination of linguistic style with sites appearance. Because these features do not rely on the actual content of pages, our approach is topic-agnostic and robust; this is in contrast to models that use content (e.g., n-grams), which must be retrained as the topics in the news shift.

(a) Fake News: Clash Daily (b) Fake News: CDP (c) Real News: Reuters

Figure 1: Web pages from unreliable and reliable new sites.

Fake News Corpora. Despite the recent research efforts focusing on fake news, there is a dearth of publicly-available datasets focusing on *web content*. Some of the public datasets relevant to fake news detection are: BS Detector[1], BuzzFeed [15], FakeNewsNet [19], NewsMediaSources [3], US-Election2016 [1] and Celebrity [14]. These datasets are limited with respect to size, the time period they cover, and variety of topics covered. Since we aim to determine the time-invariant and topic-invariant features of political fake news on the Web, these datasets are insufficient to evaluate our work.

3 TOPIC-AGNOSTIC CLASSIFICATION

In this section, we present our approach to fake news classification.

3.1 Topic-Agnostic Features

While exploring web pages containing fake news, we observed some salient topic-agnostic features. For example, the pages contain a large number of ads – this is not surprising given that providers can gain significant advertising revenue by attracting users to their web site with appealing fake news headlines [1]. Recent works have also argued that fake news articles are designed to induce inflammatory emotions in readers, and contain text patterns related to understandability that differ from mainstream news [2, 9, 14].

Figure 1 shows some examples of fake and real news pages. Fake news pages not only have a larger number of ads and polluted layouts but also have a distinctive style to their headlines, often in the form of a sensationalist slant. Additionally, besides attempting to describe the article, these headlines often contain terms designed to catch the readers' attention (e.g., "Just in", "Read this", "Breaking News"). These observations motivated us to investigate two broad classes of features: web-markup and linguistic-based (morphological, psychological and readability-related). The features are listed in Table 1 and we summarize them below.

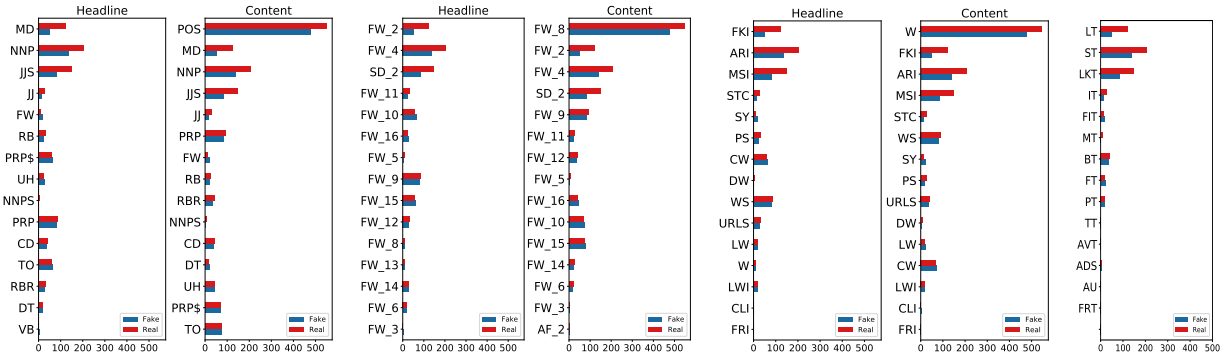
Morphological Features. This set corresponds to the frequency (word count) of morphological patterns in texts. We obtain these patterns through part-of-speech tagging, which assigns each word in a document to a category based on both its definition and context.

Psychological Features. Psychological features capture the percentage of total semantic words in texts[2]. We obtain the words'

[1] https://www.kaggle.com/mrisdal/fake-news/home
[2] http://liwc.wpengine.com/interpreting-liwc-output/

Table 1: Linguistic and web-markup features used to represent news articles.

	Abbr.	Description	Abbr.	Description	Abbr.	Description	Abbr.	Description	Abbr.	Description
Morphological Features	WDT	WH-determiner	PDT	Pre-determiner	JJ	Adjective or numeral, ordinal	VB	Verb, base form	MD	Modal auxiliary
	CD	Numeral, cardinal	VBD	Verb, past tense	VBG	Verb, present participle or gerund	VBN	Verb, past participle	RP	Particle
	DT	Determiner	NNPS	Noun, proper, plural	NN	Noun, common, singular or mass	CC	Conjunction, coordinating	WRB	Wh-adverb
	FOW	Foreign word	NNS	Noun, common, plural	TO	"To" as preposition/infinitive, superlative	WP$	WH-pronoun, possessive	JJS	Adjective, superlative
	WP	WH-pronoun	POS	Genitive marker	VBP	Verb, present tense, not 3rd singular	RBR	Adverb, comparative	NNP	Noun, proper, singular
	UH	Interjection	PRP	Pronoun, personal	VBZ	Verb, present tense, 3rd singular	RBS	Adverb superlative	PRP$	Pronoun, possessive
	RB	Adverb	JJR	Adjective, comparative	IN	Preposition or conjunction, subordinating				
Psychological Features	SO	Social (family, friend)	SD	Summary Dimensions	BP	Biological Processes (ingest, health, sexual)	AF	Affect (anger, sad, anxiety)	PC	Personal Concerns
	RL	Relativity (space, time)	FW	Function Words	TR	Time Orientation (focuspast, focuspresent)	PP	Perceptual Process	IL	Informal Language
	DR	Drives (power, risk)	PM	Punctuation Marks	OG	Other Grammar (quantifiers, interrogatives)	CP	Cognitive Processes		
Readability Features	FRI	Flesch Reading Ease	WS	Words per sentence	LW	Long words	LWI	Linsear Write	CLI	Coleman-Liau
	FKI	Flesch Kincaid Grade	CW	Capitalized words	SY	Syllables	CW	Complex words	DW	Difficult words
	MSI	McLaughlinâĂŹs SMOG	LX	Lexicon	PS	Percentage of stop words	ARI	Automated Readability	W	Words
	GFI	Gunning Fog	URL	URLs	STC	Sentences	CH	Characters		
Web-markup features	AU	Author	IT	Images (e.g., img, canvas)	ST	Semantics (e.g., article, section)	FRT	Frames (e.g., frame, frameset)	LT	Lists (e.g., ul, ol, li)
	BT	Basic (e.g., title,h1, p)	FT	Formatting (e.g., acronym)	FIT	Forms and inputs (e.g., textarea, button)	MT	Metainfo (e.g., head, meta)	TT	Tables (e.g., tbody, tfoot)
	AVT	Audio and video	LKT	Links (e.g., a, nav, link)	PT	Programming (e.g., script, object)	ADS	Advertisements		



(a) Morphological (b) Psychological (c) Readability (d) Web-markup

Figure 2: Mean frequency distribution of features per class in the PoliticalNews dataset.

semantics by using a dictionary that has lists of words that express psychological processes (personal concerns, affection, perception).

Readability Features. This set captures the ease or difficulty of comprehending a text. We obtain these features through readability scores and counting of character, words, and sentences usage.

Previous works have found that fake news often displays a divergence between the news headline and the body content [9, 20]: (i) a headline declares a piece of information to be false and the body text declares it to be true (or vice-versa); and (ii) fake news packs the main claim of the article into its title, allowing the reader to skip reading the body article, which tends to be short, repetitive, and less informative when compared with real news. These divergences between the textual pieces of news articles motivated us to apply linguistic features at different granularities: considering only the headline, only the content, and combining both. Figures 2(a), 2(b), and 2(c) show the ranking of the top-15 linguistic-based features, in PoliticalNews dataset, with the largest differences between classes considering distinct text granularities. In Figure 2, the psychological features are named alongside an index. This is because each of the psychological features represents a category that contains other features. A complete list of these features is available at http://liwc.wpengine.com/compare-dictionaries.

Web-Markup Features. These features capture patterns of the web pages' layout . The web-markup features we use include: frequency (number of occurrences) of advertisements, presence of an author name (binary value), and the frequency of distinct categories of tag groups[3]. Figure 2(d) shows the mean distribution frequency of each web-markup features in PoliticalNews dataset, within fake and real news. Note that the distributions are different, thus supporting the use of these features to distinguish news.

3.2 Feature Selection

We perform feature selection using a combination of four different methods: Shannon Entropy [12], Tree Based Rule [7], L1 Regularization [23] and Mutual Information [10]. We combine the outputs from these methods by normalizing them and applying the geometric mean to obtain a new score $r(f_i)$ which corresponds to the importance of a feature f_i:

$$r(f_i) = \sqrt[4]{SE(f_i)^{-1} \times TB(f_i) \times L1(f_i) \times MI(f_i)} \quad (1)$$

where SE refers to Shannon Entropy, TB to Tree Based Rule, L1 to L1 Regularization and MI to Mutual Information. We compute the feature importance scores and remove features with $r(f_i)$ value equal to 0. When we performed this process for features from news headlines, we found the following patterns turned out to be ineffective: FOW, IN, JJR, PRP$, TO, VBD, VBG, VBZ, WP$, MSI, CW, TT, FW (semicolon), BP (ingest), RL (time) and PC (home). When we consider features from news content, DT, PDT, RBR, RP, OG, and UH are removed. When we extract features from both

[3]https://www.w3schools.com/tags/ref_byfunc.asp

(headline + content), DT, JJS, PDT, POS, RBR, RBS, UH and WRB are eliminated. For the web-markup features, we kept IT, AVT, AU, LKT, ADS, ST and BT. The sizes of the sets of topic-agnostic features are: (1) for headlines, 137 features; (2) for content, 148; and (3) headline+content, 145. We study the effectiveness of these features in Section 4. Note that we performed the feature selection process only once, using the PoliticalNews dataset, and used the resulting features on all experiments.

3.3 Classification

We use supervised learning to classify news based on the selected features. Formally, the resulting classifier corresponds to a function that takes as input the TAG features of a web page $x \in \mathbb{R}^d$ and produces an output $\hat{y} \in C$, where C is the set of all categories. Here we use two categories: fake and real news. We experimented with three different learning methods: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest (RF). We use SVM classifier with linear kernel and cost equal to 0.1. The other methods are setup with default parameters of the scikit-learn[4] library.

4 EXPERIMENTAL EVALUATION

In this section, we describe the evaluation that we applied to assess how effective our approach is at identifying fake news.

Corpus. As discussed in Section 2, existing public datasets are limited. So, we built a new dataset that we call ***PoliticalNews***[5], by compiling a list of known reliable and unreliable sites and crawling them. As seeds for the crawl, we used web sites coming from Politifact[6], BuzzFeed [15], and OpenSources.co[7] as unreliable news sites and 242 sites from the most visited news sites *Alexa's top 500 news sites*[8] as reliable sites. Since the last one is based just on popularity, we manually selected the web sites focused on political news. One challenge in constructing such a large dataset is the impracticality of individually labeling all the articles. Our approach was to project the domain-level ground truth onto the content collected from those domains. We collected over 1.6 million web pages published between 2013 and 2018. Then, we post-processed the data removing non-political pages using a Naïve Bayes model combined with TF-IDF feature representation, and as a training set, we use a publicly-available corpus[9]. Furthermore, to ensure a balanced distribution of web pages sites for each year, we sampled 32 pages from each site for each of the years we have crawl data for. The result of this balancing is a dataset of 14,240 news pages with 7,136 pages coming from 79 unreliable sites, and another 7,104 coming from 58 reliable sites. The motivation for this balancing is to avoid the problem of overfitting (since real news is orders of magnitude more prevalent than fake news). Additionally, ensuring balance across the years covered allows us to evaluate the effectiveness of different approaches as news topics change over time.

To demonstrate the robustness of the TAG model, we also evaluate our approach over two additional datasets that have been used in previous works: ***Celebrity*** [14] and ***US-Election2016*** [1]. We note that these datasets contain only the text of the web pages. Since we need to extract web-markup features, we fetched the original web page source from the Web. For sites and pages that were no longer available, we retrieved versions from the Web Archive. In addition, duplicated articles were removed. As a result, the versions of the Celebrity and US-Election2016 datasets[10] used in our experiments contain 479 and 691 news articles, respectively.

Experimental Setup. We use the NLTK [5] library part-of-speech tagger to compute *Morphological Features*. To extract the *Psychological Features*, we use the Linguistic Inquiry and Word Count software (LIWC, Version 1.3.1 2015) [13]. To compute *Readability Features*, we use a Textstat[11] library, and finally, to extract the *Web-Markup Features* from web pages, we use BeautifulSoup[12] and Newspaper.[13]

We compare the TAG classifier with the Fake News Detector (FNDetector) presented in [14]. This model represents documents by using four sets of linguistic features: n-grams (unigrams + bigrams encoded by TF-IDF values), psychological, readability and syntactical features (production rules of context-free grammars encoded by TF-IDF values). The psychological and readability sets are the same as we use (see Table 1). We selected the FNDetector for two main reasons: like our approach, they focus on the automatic identification of fake content in online news; and because their classifier attained high accuracy using content-based features, it is a suitable baseline for our TAG classifier. To better understand the contribution of individual features, we created multiple baseline classifiers using different feature combinations. We used a linear SVM classifier and conducted our evaluations using five-fold cross-validation with accuracy as the performance metric.

4.1 Effectiveness of Different Features

We evaluate the performance of models trained with different combinations of feature sets (separately and jointly). In addition, to assess the performance of the classifiers using different representations for a news article, each experiment was performed for the headline (H), content (C) and the combination of both (HC).

Table 2 shows the accuracy obtained for the different TAG feature sets over the three datasets. Note that combining features often leads to the highest accuracies. For example, for political news, the highest accuracies are obtained by the configuration that combines **LIWC (L) - NLTK (N) - readability (R) - webmarkup (W)**, which attains 0.86 accuracy for US-Election2016 and 0.83 for PoliticalNews. For the Celebrity data, the best configuration is **LIWC (L) - readability (R) - webmarkup (W)** which attains 0.78 accuracy. If we further examine the results from LIWC, we can see that combining LIWC with other features often leads to higher accuracies: for all datasets, the accuracy gains vary between 0.012 and 0.21. This reinforces previous findings [2, 14] that psychological factors play an important role in the disinformation ecosystem. The web-markup features also lead to improvements, in particular when compared against readability for the PoliticalNews corpus.

[4] http://scikit-learn.org/
[5] https://osf.io/ez5q4/
[6] https://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/
[7] http://www.opensources.co
[8] https://www.alexa.com/topsites/category/News
[9] https://www.kaggle.com/rmisra/news-category-dataset/home
[10] https://osf.io/qj86g/
[11] http://pypi.python.org/pypi/textstat/
[12] https://www.crummy.com/software/BeautifulSoup
[13] https://newspaper.readthedocs.io/en/latest

Table 2: Accuracy results for models that use different set of topic-agnostic features – where L is LIWC, N is NLTK, R is readability, and W is webmarkup features – over three different datasets: Celebrity, US-Election2016, and PoliticalNews. The best accuracies for each feature set are bold; the best accuracies for each news article's representations (H, C and HC) are underlined.

Dataset	Celebrity			US-Election2016			PoliticalNews		
Features	H	C	HC	H	C	HC	H	C	HC
W	**0.68**	**0.68**	**0.68**	**0.65**	**0.65**	**0.65**	**0.71**	**0.71**	**0.71**
L	0.69	**0.73**	**0.73**	0.77	0.81	**0.83**	0.71	0.75	**0.76**
N	0.58	**0.68**	0.66	**0.81**	0.75	0.76	**0.77**	0.66	0.67
R	0.57	**0.62**	0.57	**0.75**	0.73	0.73	**0.69**	0.62	0.64
L-R	0.65	**0.76**	0.74	0.79	0.82	**0.83**	0.74	0.75	**0.76**
N-R	0.65	**0.68**	0.67	**0.83**	0.78	0.78	**0.78**	0.71	0.72
N-W	0.68	**0.72**	**0.72**	0.79	0.79	**0.80**	**0.81**	0.79	0.79
L-W	0.70	**0.77**	0.75	0.79	0.82	**0.83**	0.78	**0.81**	**0.81**
R-W	0.67	**0.72**	0.67	**0.80**	0.76	0.76	**0.77**	0.76	0.76
N-R-W	0.67	**0.72**	0.71	**0.83**	0.79	0.80	**0.81**	0.78	0.79
L-R-W	0.71	**0.78**	0.71	0.79	0.83	**0.85**	0.80	0.80	**0.81**
L-N-R-W	**0.73**	**0.73**	0.71	0.83	0.82	**0.86**	**0.83**	0.81	0.82

Table 3: Classification results (accuracies) for three datasets.

Dataset	Celebrity	US-Election2016	PoliticalNews
FNDetector	0.73	0.81	0.76
TAG Model	**0.78**	**0.86**	**0.83**

When we consider the features from different article parts (H, C, and HC), the distribution of accuracies for US-Election2016 and PoliticalNews are similar. This suggests that the headlines contain useful information that allows the identification of political fake news. Thus, either headlines or content or both can be used for classification. On the other hand, for the Celebrity dataset, we can clearly see that the classifier achieves slightly better results using the content of the articles. Consider the following two examples from this dataset: (1) a real news, where the headline is "*Stop Right Now! The Spice Girls Might Be Planning A Reunion*" and the content is "*There are certain things that people just shouldn't joke about. A potential Spice Girls reunion with all five members is one of them. Earlier this morning...*", and (2) a fake news, where the headline is "*Taylor Swift Goes Naked in '...Ready for It?' Watch!*" and the content is "*Nope, definitely not ready for this! Taylor Swift released a 15-second teaser for the music video for her new song '... Ready for It?' on Monday...*". The previous example in the Celebrity dataset shows a concrete example where headlines (H) for fake and real news have similar linguistic features (e.g., use of capitalized words, punctuation), but content (C) is more discriminative in terms of linguistic features and readability. Furthermore, the web-markup features are effective in this scenario, since they make it possible to detect a clear difference between fake and real news based on web page characteristics (e.g., number of ads, links, and images).

We also compare the best TAG feature set combination for each dataset, identified during our previous experiment, with the baseline. We use a linear SVM and five-fold cross-validation with accuracy as metric. As shown in Table 3, the classifier based on TAG

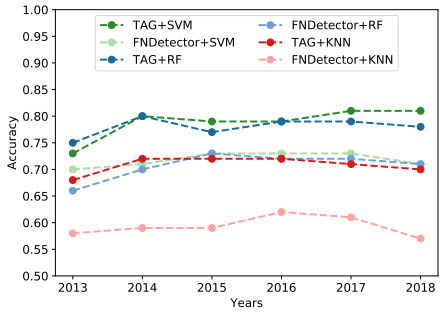
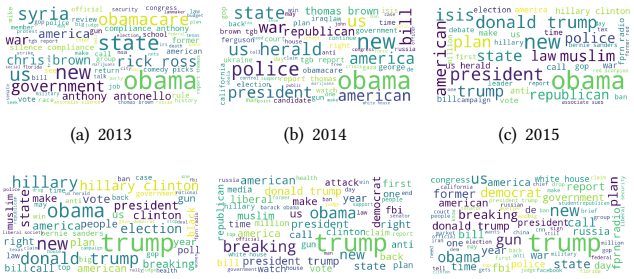


Figure 3: Effectiveness of SVM, KNN and RF using baseline (FNDetector) and TAG model in different time windows.



(a) 2013 (b) 2014 (c) 2015

(d) 2016 (e) 2017 (f) 2018

Figure 4: Tag clouds of the most frequent terms in web pages.

features outperforms the *FNDetector* for all datasets, indicating that the task of fake news classification can be effectively accomplished using topic-agnostic features. Furthermore, it is important to note that *FNDetector* features uses 3 orders of magnitude more features (~**800,000**) than our approach (~**160** features). The number of features has important implications for the processing time, maintaining the model, and explainability.

4.2 Effectiveness over Time

To study the behavior of our approach as news topics change over time, we split PoliticalNews dataset into six time windows (sub-datasets) corresponding to each year from 2013 to 2018. We then experimented with multiple classifier configurations: each configuration uses one sub-dataset for training and the others (one at a time) for testing. For example, we use pages published in 2013 to construct a classifier C_{2013} and use C_{2013} to classify pages in the sub-datasets from 2014 through 2018. Note that each sub-dataset is associated with 5 results. For this experiment, our TAG model uses the set of features NLTK, LIWC, readability and web-markup.

Figure 3 shows the average accuracies for each sub-dataset over the 5 test sets. Our TAG model performs better than *FNDetector* for all the time windows. The relatively lower accuracy values obtained by *FNDetector* can be explained by its dependence on the contents. You can observe the topic changes in Figure 4, which shows tag clouds (built over our training data using n-grams frequencies) summarizing the news in each year. Note that even though the tag clouds share some keywords (e.g., "Obama" and "Trump"), they appear at different frequencies and cover distinct subjects.

The content-based model essentially learns how to detect fake news for a specific time and topics. Note, for example, the difference between the topics in 2013 and in 2015. In 2013, "Trump" was

Table 4: Cross-domain results (accuracies) between models.

Training	Test	Classifier	Accuracy	
			FNDetector	TAG Model
Celebrity	US-Election 2016	SVM	0.59	**0.70**
		KNN	0.59	**0.64**
		RF	0.56	**0.64**
US-Election 2016	Celebrity	SVM	0.59	**0.63**
		KNN	0.56	**0.60**
		RF	0.51	**0.60**

not yet involved in politics (see Figure 4(a)), but he starts to appear in political news in 2015, and more frequently since the 2016 elections as shown in Figure 4(e) and Figure 4(f). Also note the differences between 2016 and 2017: In 2016, political news were centered around the campaign, parties, voters, etc., and we see terms such as "Hillary", "Clinton", "vote", "Trump" and "Obama". But in 2017 (after the elections), the terms "Hillary", "Clinton" decrease and terms like "breaking", "attack", "Russia", "FBI" start to appear in the political discourse. This indicates that, to be effective, content-based models must be constantly retrained. This is both costly and error-prone.

4.3 Effectiveness for Different Domains

We also evaluated the generalizability of our approach when the training and testing sets are drawn from entirely different topic domains. We ran two experiments: in the first, we use Celebrity as a training dataset and US-Election2016 as testing dataset; and for the second experiment, the other way around. We use all the TAG feature sets identified during our previous experiments (N, L, R and W). We considered three classifiers – SVM, KNN and RF; and used five-fold cross-validation. The results in Table 4 show that our topic-agnostic approach attains accuracies that are substantially higher than those of FNDetector. While our original goal in this work was to design a classifier that is able to distinguish fake and real political news as they evolve over time, these results show that the approach is promising for topic domains beyond political news.

5 CONCLUSIONS & FUTURE WORK

We presented a new approach to detect fake news web pages which uses topic-agnostic features. Through a detailed experimental evaluation, we showed that our approach accurately classifies not only political news as topics evolves over time but also news from different domains, outperforming content-based approaches while using significantly fewer features and requiring no frequent re-training. Our results suggest that topic-agnostic features are effective for distinguishing between fake and real news, and that robust classifiers can be constructed that enable the timely discovery of fake news articles. We have also created a new corpus of over 14,000 political news pages drawn from 137 sites and spanning 6 years. To the best of our knowledge, this is the first of its kind in terms of size, focus on the Web, and inclusion of HTML markup information.

There are many directions we plan to pursue in future work. We will explore further improvements to the topic-agnostic model by considering additional features, for example, user engagement and network structure. We would also like to experiment with different strategies to expand our fake news corpus, including the use of social media to search for previously unknown sites, and by using the top-agnostic classifier in conjunction with a focused crawler to discover new sites as they are created.

ACKNOWLEDGMENTS

This work was partially supported by the DARPA MEMEX and D3M programs, NSF award CNS-1229185, and the Brazilian National Council for the Improvement of Higher Education (CAPES) under grant 88887.130294/2017-00. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31 (2017), 211–36.
[2] Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism* 6 (2018), 154–175.
[3] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proc. of the Conf. on Empirical Methods in NLP*. 3528–3539.
[4] Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A System to Unmask Propaganda in Online News. In *Proc. of the 33th AAAI Conf. on Artificial Intelligence*.
[5] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proc. of the Association for Computational Linguistics on Interactive poster and demonstration sessions*. 31.
[6] James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility Assessment in the News: Do we need to read?. In *Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining*. 799–800.
[7] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63 (2006), 3–42.
[8] Jeffrey Gottfried and Elisa Shearer. 2016. News Use across Social Media Platforms 2016. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016.
[9] Benjamin D Horne and Sibel Adali. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proc. of the 2nd Intn'l Workshop on News and Public Opinion*.
[10] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69 (2004), 066138.
[11] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* (2018).
[12] Annick Lesne. 2014. Shannon Entropy: A Rigorous Notion at The Crossroads Between Probability, Information Theory, Dynamical Systems and Statistical Physics. *Mathematical Structures in Computer Science* 24 (2014), 240–311.
[13] James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. 10.15781/T29G6Z.
[14] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proc. of Int'l Conf. on Computational Linguistics*. 3391–3401.
[15] Scott Pham. [n. d.]. Analysis of fake news sites and viral posts, 2016 vs. 2017. https://github.com/BuzzFeedNews/2017-12-fake-news-top-50.
[16] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A Stylometric Inquiry into Hyperpartisan and Fake News. *CoRR* abs/1702.05638 (2017).
[17] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proc. of the Conf. on Empirical Methods in NLP*. 2931–2937.
[18] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19 (2017), 22–36.
[19] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting Tri-Relationship for Fake News Detection. *CoRR* abs1712.07709 (2017).
[20] Craig Silverman. 2015. Lies, damn lies, and viral content: How news websites spread (and Debunk) online rumors, unverified claims and misinformation. *Tow Center for Digital Journalism* 168 (2015).
[21] Craig Silverman and Jeremy Singer-Vine. 2016. Most Americans Who See Fake News Believe It, New Survey Says. BuzzFeed News.
[22] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proc. of the 11th Int'l Conf. on Web Search and Data Mining*. 637–645.
[23] Peng Zhao and Bin Yu. 2006. On model selection consistency of Lasso. *Journal of Machine learning research* 7 (2006), 2541–2563.

Differences in Health News from Reliable and Unreliable Media

Sameer Dhoju
University of Mississippi
sdhoju@go.olemiss.edu

Md Main Uddin Rony
University of Mississippi
mrony@go.olemiss.edu

Muhammad Ashad Kabir
Charles Sturt University
akabir@csu.edu.au

Naeemul Hassan
University of Mississippi
nhassan@olemiss.edu

ABSTRACT

The spread of 'fake' health news is a big problem with even bigger consequences. In this study, we examine a collection of health-related news articles published by reliable and unreliable media outlets. Our analysis shows that there are structural, topical, and semantic patterns which are different in contents from reliable and unreliable media outlets. Using machine learning, we leverage these patterns and build classification models to identify the source (reliable or unreliable) of a health-related news article. Our model can predict the source of an article with an F-measure of 96%. We argue that the findings from this study will be useful for combating the health disinformation problem.

CCS CONCEPTS

• **Applied computing → Health informatics**.

ACM Reference Format:
Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in Health News from Reliable and Unreliable Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316741

1 INTRODUCTION

Of the 20 most-shared articles on Facebook in 2016 with the word "cancer" in the headline, more than half the reports were discredited by doctors and health authorities [7]. The spread of health-related hoaxes is not new. However, the advent of Internet, social networking sites (SNS), and click-through-rate (CTR)-based pay policies have made it possible to create hoaxes/"fake news", published in a larger scale and reach to a broader audience with a higher speed than ever [16]. Consequences of misleading or erroneous health news can be very critical. Believing health misinformation may lead to a hazardous health condition. Houston reported a measles outbreak in Europe due to lower immunization rate which experts believed was the result of anti-vaccination campaigns caused by a false news about MMR vaccine [14]. Moreover, misinformation can spoil the credibility of the health-care providers and create a lack of trust in taking medicine, food, and vaccines.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316741

Recently, researchers have started to address the fake news problem in general [21, 32]. However, health disinformation is a relatively unexplored area. According to a report from Pew Research Center [8], 72% of adult internet users search online for information about a range of health issues. So, it is important to ensure that the health information which is available online is accurate and of good quality. There are some authoritative and reliable entities such as National Institutes of Health (NIH) [1] or *Health On the Net* [2] which provide high-quality health information. Also, there are some fact-checking sites such as Snopes.com [3] and Quackwatch.org [4] that regularly debunk health and medical related misinformation. Nonetheless, these sites are incapable of busting the deluge of health disinformation continuously produced by unreliable health information outlets (e.g., RealFarmacy.com, Health Nut News). Moreover, the bots in social networks significantly promote unsubstantiated health-related claims [9]. Researchers have tried developing automated health hoax detection techniques but had limited success due to several reasons such as small training data size and lack of consciousness of users [11, 12, 20, 35].

The objective of this project is **(i)** to develop a health-oriented news dataset that covers both reliable and unreliable media outlets, **(ii)** to identify discriminating features that can potentially separate a reliable health news from an unreliable health news by leveraging a large-scale dataset, and **iii)** use the dataset and the discriminating features for building a classifier. We examine how reliable media and unreliable media outlets conduct health journalism. First, we prepare a large dataset of health-related news articles which were produced and published by a set of reliable media outlets and unreliable media outlets. Then, using a systematic content analysis, we identify the features which separate a reliable outlet sourced health article from an unreliable sourced one. These features incorporate the structural, topical, and semantic differences in health articles from these outlets. For instance, our structural analysis finds that the unreliable media outlets use clickbaity headlines in their health-related news significantly more than what reliable outlets do. Our topical analysis finds that while the reliable outlets discuss "cancer" along with research and studies, in the unreliable outlets "cancer" is associated with autism and vaccination. The semantic analysis shows that on average a health news from reliable media contains more reference quotes than an average unreliable sourced health

[1] https://www.nih.gov/
[2] https://www.hon.ch/en/
[3] https://www.snopes.com/
[4] http://www.quackwatch.org/

news. We argue that these features can be critical in understanding health misinformation and designing systems to combat such disinformation.

To justify this, we developed a classifier to distinguish unreliable media sourced health news from reliable articles. Along with the word vectors built from the body of the articles we utilized these features from our analysis to build the classification model. Experiment results demonstrated promising accuracy of our approach. Further, we identified and analyzed the most-important features behind the model. Our analysis shows that 4 of our engineered features were among the 20 most-important features.

2 RELATED WORK

Fake news is an emerging topic which draws continuous attractions of the researchers. There are already some approaches for detecting fake news computationally. Potthast et al. use lexical and syntactic features, capture specific writing styles to build their fake news detection model [27]. Rubin et al. also capture manipulation in writing styles to identify fake news [2]. Apart from writing styles, Gupta et al. use visual features to identify fake images that are created intentionally [13]. [23, 39] employ external knowledge bases to verify the claims in the news contents. All the mentioned approaches explored fake news identification problem from a general perspective. None of them examined domain specific features for example features related to health misinformation which could be instrumental to identify fake news more effectively.

There has been extensive work on how scientific medical research outcomes should be disseminated to general people by following health journalism protocols [4, 5, 18, 31, 33]. For instance, Lopes et al. suggest that it is necessary to integrate journalism studies, strategic communication concepts, and health professional knowledge to successfully disseminate professional findings. Some researchers particularly focused on the spread of health misinformation in social media [22]. For example, Chou et al. discuss the gap in understanding the effect of health misinformation, challenges in developing a framework for research and practice on social media [3]. [11] analyzes Zika [5] related misinformation in Twitter. In particular, it shows that tracking health misinformation in social media is not trivial, and requires some expert supervision. It exploited crowdsource to annotate a collection of Tweets and used the annotated data to build a rumor classification model. One limitation of this work is that the used dataset is too small (6 rumors) to make a general conclusion. Moreover, it didn't consider the features in the actual news articles unlike us. [12] examines the individuals on social media that are posting questionable health-related information, and in particular promoting cancer treatments which have been shown to be ineffective. It develops a feature based supervised classification model to automatically identify users who are comparatively more susceptible to health misinformation. There are other works which focus on automatically identifying health misinformation. For example, [19] developed a classifier to detect misinformative posts in health forums. One of the limitations of this work is that the training data is only labeled by two individuals. Researchers have also worked on building tools that can help a user to easily consume health information. [20] developed the "VAC Medi+board", an interactive visualization platform integrating Twitter data and news coverage from a reliable source called MediSys[6]. It covers public debate related to vaccines and helps users to easily browse health information on a certain vaccine-related topic.

Our study significantly differs from these already existing researches. Instead of depending on a small sample of health hoaxes like some of the existing works, we take a different approach and focus on the source outlets. This gives us the benefit of investigating with a larger dataset. We investigate the journalistic practice of reliable and unreliable health outlets, an area which has not been studied to the best of our knowledge.

3 DATA PREPARATION

For investigating how reliable media outlets and unreliable outlets portray health information, we need a reasonably sized collection of health-related news articles from these two sides. Unfortunately, there is not an available dataset which is of adequate size. For this reason, we prepare a dataset of about 30, 000 health-related news articles disseminated by reliable or unreliable outlets within the years 2015 – 2018. Below, we describe the preparation process in detail.

3.1 Media Outlet Selection

The first challenge is to identify reliable and unreliable outlets. The matter of reliability is subjective. We decided to consider the outlets which have been cross-checked as reliable or unreliable by credible sources.

3.1.1 Reliable Media. We identified 29 reliable media outlets from three sources– **i)** 11 of them are certified by the Health On the Net [25], a non-profit organization that promotes transparent and reliable health information online. It is officially related with the World Health Organization (WHO) [36]. **ii)** 8 from U.S. government's health-related centers and institutions (e.g., CDC, NIH, NCBI), and **iii)** 10 from the most circulated broadcast [30] mainstream media outlets (e.g., CNN, NBC). Note, the mainstream outlets generally have a separate section for health information (e.g., https://www.cnn.com/health). As our goal is to collect health-related news, we restricted ourselves to their health portals only.

3.1.2 Unreliable Media. Dr. Melissa Zimdars, a communication and media expert, prepared a list of false, misleading, clickbaity, and satirical media outlets [38, 40]. Similar lists are also maintained by Wikipedia [37] and informationisbeautiful.net [15]. We identified 6 media outlets which primarily spread health-related misinformation and are present in these lists. Another source for identifying unreliable outlets is *Snopes.com*, a popular hoax-debunking website that fact-checks news of different domains including health. We followed the health or medical hoaxes debunked by *Snopes.com* and identified 14 media outlets which sourced those hoaxes. In total, we identified 20 unreliable outlets. Table 1 lists the Facebook page ids of all the reliable and unreliable outlets that have been used in this study.

[5] https://en.wikipedia.org/wiki/Zika_virus

[6] http://medisys.newsbrief.eu

Table 1: List of Facebook page ids of the reliable and unreliable outlets. Some of them are unavailable now.

Reliable	everydayhealth, WebMD, statnews, AmericanHeart, BBCLifestyleHealth, CBSHealth, FoxNewsHealth, WellNYT, latimesscience, tampabaytimeshealth, philly.comhealth, AmericanHeart, AmericanCancerSociety, HHS, CNNHealth, cancer.gov, FDA, mplus.gov, NHLBI, kidshealthparents, ahrq.gov, healthadvocateinc, HealthCentral, eMedicineHealth, C4YWH, BabyCenter, MayoClinic, MedicineNet, healthline
Unreliable	liveahealth, healthexpertgroup, healthysolo, organichealthcorner, justhealthylifestyle1, REALfarmacyCOM, thetruthaboutcancer, BookforHealthyLife, viralstories.bm, justhealthyway, thereadersfile, pinoyhomeremedies, onlygenuinehealth, greatremediesgreathealth, HealthRanger, thefoodbabe, AgeofAutism, HealthNutNews, consciouslifenews, HealthImpactNews

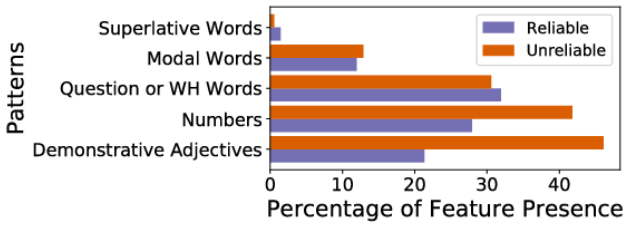


Figure 1: Distribution of clickbait patterns

3.2 Data Collection

The next challenge is to gather news articles published by the selected outlets. We identified the official Facebook pages of each of the 49 media outlets and collected all the link-posts [7] shared by the outlets within January 1, 2015 and April 2, 2018 [8] using Facebook Graph API. For each post, we gathered the corresponding news article link, the status message, and the posting date.

3.2.1 News Article Scraping. We used a Python package named Newspaper3k [9] to gather the news article related data. Given a news article link, this package provides the headline, body, author name (if present), and publish date of the article. It also provides the visual elements (image, video) used in an article. In total, we collected data for 29, 047 articles from reliable outlets and 15, 017 from unreliable outlets.

3.2.2 Filtering non-Health News Articles. Even though we restricted ourselves to health-related outlets, we observed that the outlets also published or shared non-health (e.g., sports, entertainment, weather) news. We removed these non-health articles from our dataset and only kept *health*, *food & drink*, or *fitness & beauty* related articles. Specifically, for each news article, we used the document categorization service provided by Google Cloud Natural Language API [10] to determine its topic. If an article doesn't belong to one of the three above mentioned topics, it is filtered out. This step reduced the dataset size to 27, 589; 18, 436 from reliable outlets and 9, 153 from unreliable outlets. We used this health-related dataset only in all the experiments of this paper.

4 CONTENT ANALYSIS

Using this dataset, we conduct content analysis to examine structural, topical, and semantic differences in health news from reliable and unreliable outlets.

4.1 Structural Difference

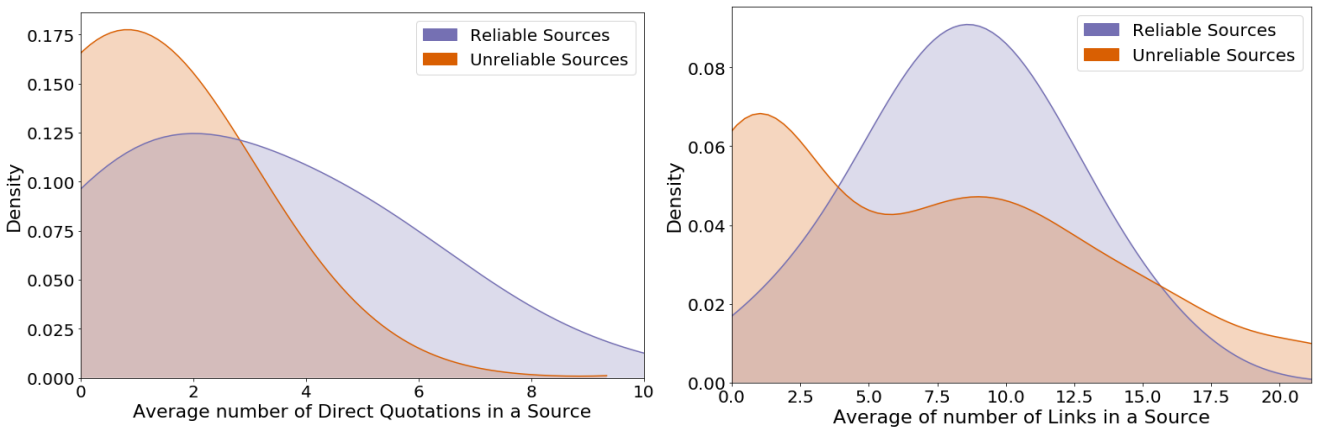
The headline is a key element of a news article. According to a study done by American Press Institute and the Associated Press [17], only 4 out of 10 Americans read beyond the headline. So, it is important to understand how reliable and unreliable outlets construct the headlines of their health-related news. According to to [1], a long headline results in significantly higher click-through-rate (CTR) than a short headline does. We observe that the average headline length of an article from reliable outlets and an article from unreliable outlets is 8.56 words and 12.13 words, respectively. So, on average, an unreliable outlet's headline has a higher chance of receiving more clicks or attention than a reliable outlet's headline. To further investigate this, we examine the *clickbaityness* of the headlines. The term clickbait refers to a form of web content (headline, image, thumbnail, etc.) that employs writing formulas, linguistic techniques, and suspense creating visual elements to trick readers into clicking links, but does not deliver on its promises [10]. Chen et al. [2] reported that clickbait usage is a common pattern in false news articles. We investigate to what extent the reliable and unreliable outlets use clickbait headlines in their health articles. For each article headline, we test whether it is a clickbait or not using two supervised clickbait detection models– a sub-word embedding based deep learning model [28] and a feature engineering based Multinomial Naive Bayes model [24]. Agreement between these models was measured as 0.44 using Cohen's κ. We mark a headline as a clickbait if both models labeled it as clickbait. We observe, 27.29% (5,031 out of 18,436) of the headlines from reliable outlets are click bait. In unreliable outlets, the percentage is significantly higher, 40.03% (3,664 out of 9,153). So, it is evident that the unreliable outlets use more click baits than reliable outlets in their health journalism.

We further investigate the linguistic patterns used in the clickbait headlines. In particular, we analyze the presence of some common patterns which are generally employed in clickbait according to [1, 26]. The patterns are-

- Presence of demonstrative adjectives (e.g., this, these, that)
- Presence of numbers (e.g., 10, ten)
- Presence of modal words (e.g., must, should, could, can)
- Presence of question or WH words (e.g., what, who, how)
- Presence of superlative words (e.g., best, worst, never)

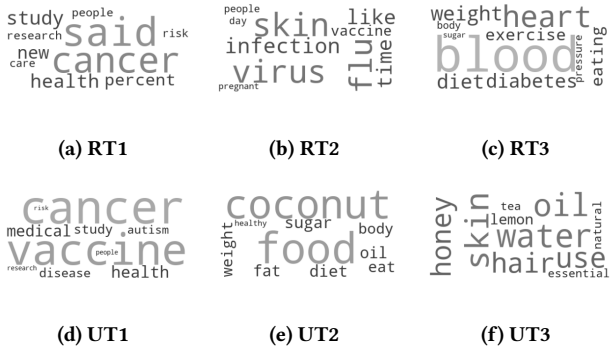
Figure 1 shows the distribution of these patterns among the clickbait headlines of reliable and unreliable outlets. Note, one headline may contain more than one pattern. For example, this headline *"Are these the worst 9 diseases in the world?"* contains four of the above patterns. This is the reason why summation of the percentages isn't equal to one. We see that unreliable outlets use demonstrative adjective and numbers significantly more compared to the reliable outlets.

[7] Facebook allows posting status, pictures, videos, events, links, etc. We collected the link type posts only.
[8] After that, Facebook limited access to pages as a result of the Cambridge Analytica incident.
[9] https://newspaper.readthedocs.io/en/latest/
[10] https://cloud.google.com/natural-language/



(a) Quotation (b) Link

Figure 2: Distribution of average number of quotation/link per article from reliable and unreliable outlets.



(a) RT1 (b) RT2 (c) RT3

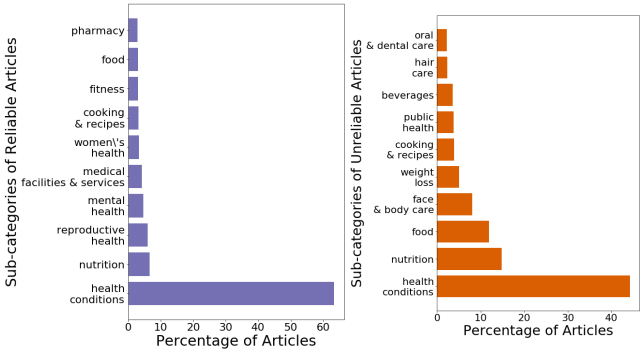
(d) UT1 (e) UT2 (f) UT3

Figure 3: Topic modeling ($k = 3$) of articles from reliable outlets (top, denoted as RT) and from unreliable outlets (bottom, denoted as UT).

4.2 Topical Difference

All the articles which we examined are health-related. However, the health domain is considerably broad and it covers many topics. We hypothesize that there are differences between the health topics which are discussed in reliable outlets and in unreliable outlets. To test that, we conduct an unsupervised and a supervised analysis.

4.2.1 Topic Modeling. We use *Latent Dirichlet Allocation(LDA)* algorithm to model the topics in the news articles. The number of topics, k, was set as 3. Figure 3 shows three topics for each of the outlet categories. Each topic is modeled by the top-10 important words in that topic. The font size of words is proportional to the importance. Figure 3a and 3d indicate that "cancer" is a common topic in reliable and unreliable outlets. Although, the words *study*, *said*, *percent*, *research*, and their font sizes in Figure 3a indicate that the topic "cancer" is associated with research studies, facts, and references in reliable outlets. On the contrary, unreliable outlets have the words *vaccine*, *autism*, and *risk* in Figure 3d which suggests the discussion regarding how vaccines put people under autism and cancer risk, an unsubstantiated claim, generally propagated by unreliable media [11, 12]. Figure 3e and 3f suggest the discussions about weight loss, skin, and hair care products (e.g., essential oil, lemon). Topics in Figure 3b and 3c discuss mostly flu, virus, skin infection, exercise, diabetes and so on.



(a) Reliable (b) Unreliable

Figure 4: Top-10 topics in reliable and unreliable outlets.

4.2.2 Topic Categorization. In addition to topic modeling, we categorically analyze the articles' topics using Google Cloud Natural Language API [13]. Figure 4 shows the top-10 topics in the reliable and unreliable outlets. In the case of reliable, the distribution is significantly dominated by *health condition*. On the other hand, in the case of unreliable outlets, percentages of *nutrition* and *food* are noticeable. Only 4 of the 10 categories are common in two outlet groups. Unreliable topics have *weight loss*, *hair care*, *face & body care*. This finding supports our claim from topic modeling analysis.

[11] https://www.webmd.com/brain/autism/do-vaccines-cause-autism
[12] https://www.skepticalraptor.com/skepticalraptorblog.php/polio-vaccine-causes-cancer-myth/
[13] https://cloud.google.com/natural-language/

4.3 Semantic Difference

We analyze what efforts the outlets make for a logical and meaningful health news. Specifically, we consider to what extent the outlets use quotations and hyperlinks. Use of quotation and hyperlinks in a news article is associated with credibility [6, 34]. Presence of quotation and hyperlinks indicates that an article is logically constructed and supported with credible factual information.

4.3.1 Quotation. We use the Stanford QuoteAnnotator [14] to identify the quotations from a news article. Figure 2a shows density plots of the number of quotations per article for reliable and unreliable outlets. We observe that unreliable outlets use less number of quotations compared to reliable outlets. We find that the average number of quotations per article is 1 and 3 in unreliable and reliable outlets, respectively. This suggests that the reliable outlet sources articles are more credible and unreliable outlets are less credible.

4.3.2 Hyperlink. We examine the use of the hyperlink in the articles. On average, a reliable outlet sourced article contains 8.4 hyperlinks and an unreliable outlet sourced article contains 6.8 hyperlinks. Figure 2b shows density plots of the number of links per article for reliable and unreliable outlets. The peaks indicate that most of the articles from reliable outlets have close to 8 (median) hyperlinks. On the other hand, most of the unreliable outlet articles have less than 2 hyperlinks. This analysis again suggests that the reliable sourced articles are more credible than unreliable outlet articles.

5 SOURCE CLASSIFICATION

In this section, we describe our classification model design process. The model leverages the above mentioned features and uses other generic text features.

5.1 Feature Extraction

We extract two types of features from the news articles. They are described below.

Word (W): We used n-grams ($n = 1, 2$) that are present in the body of the articles to build *tf-idf* word features. We discarded rare n-grams that appeared in less than five articles. We also set the maximum number of n-grams to 5,000 to avoid over-fitting. These 5,000 n-grams are the most-frequent in the whole corpus. We didn't apply stemming. However, we removed the English stop words. We carefully removed any mention of the source of the article from the features. We also removed the author names. These steps ensure that our model is not learning news outlet or author categorization.

Extracted Features (EF): This set of features were extracted from the news articles following the section 4. The features are- *headline length*, *5 linguistic patterns from headline*, *clickbaitiness*, *number of direct quotes*, and *number of hyperlinks*. In addition, we considered the presence of social media mention as a source in the news article as a feature. To identify whether an article contains any social media mention, we applied the approach described in [29].

[14] https://stanfordnlp.github.io/CoreNLP/quote.html

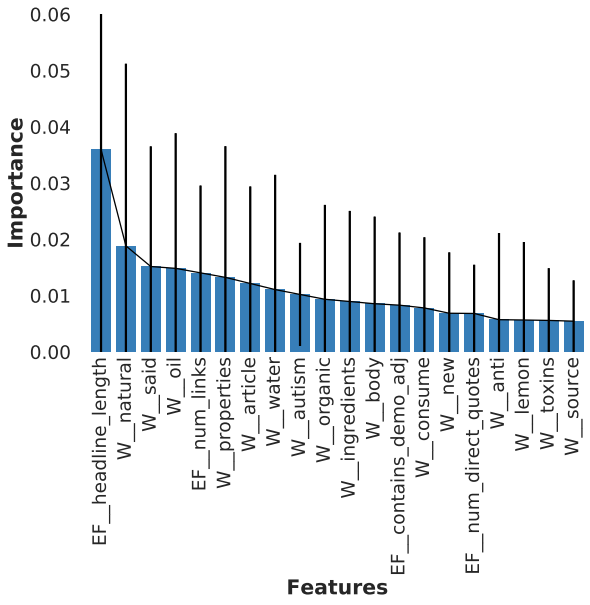


Figure 5: Feature Importance

5.2 Feature Importance

To compare the effectiveness of the features, we measured the importance of each feature. We trained a random forest classifier for which we used 50 estimators to measure the importance of features in constructing each decision tree. The overall importance of a feature is its average importance over all the trees. Figure 5 shows the importance of the 20 most-important features in the forest. The black solid lines indicate the standard deviations of the importance values. To distinguish between ngram features (W) and extracted features (EF), we use their short form as prefixes. We find that 4 out of 10 EF's make to top 20 including the top spot. Headline length, number of direct quotes, and hyperlinks proves to be very important features. It is unsurprising that words such as autism, toxic, lemon, mixture are present in the most important features.

5.3 Classification

Our dataset contains 27,589 health related articles. We preformed 5-fold cross-validation using several classical machine learning methods, including Multinomial Naive Bayes, Linear Support Vector Classifier (*SVC*), Random Forest, and Logistic Regression. Among them, Multinomial Naive Bayes and *SVC* were set with default parameters. We tried Logistic Regression and Random Forest with random state 0. For Random Forest, the number of estimators and the maximum depth were set to 200 and 3 respectively. We conducted experiments with three combinations of features - Words or n-grams (*W*), Extarcted Features (*EF*), and Word + Extracted Features (*W + EF*). For all the combinations, *SVC* outperformed other models. So, we reported here the performance of *SVC* only. Table 2 shows the performance in terms of precision, recall, and f-measure. Feature set, *W* outperforms *EF* by a large margin. The combination of both feature sets (*W+EF*) improves the overall performance, 96% average (macro) F-measure.

Table 2: Classification Report for different combination of features

Features	Labels	Precision	Recall	F-1
Word (W)	Unreliable	0.94	0.92	0.93
	Reliable	0.96	**0.97**	**0.97**
	Macro-Avg	0.95	**0.95**	0.95
Extracted Features (EF)	Unreliable	0.76	0.47	0.58
	Reliable	0.78	0.93	0.85
	Macro-Avg	0.77	0.70	0.72
W + EF	Unreliable	**0.95**	**0.93**	**0.94**
	Reliable	**0.97**	**0.97**	**0.97**
	Macro-Avg	**0.96**	**0.95**	**0.96**

6 CONCLUSION AND FUTURE WORK

In this paper, we closely looked at structural, topical, and semantic differences between articles from reliable and unreliable outlets. Our findings reconfirm some of the existing claims such as unreliable outlets use clickbaity headlines to catch the attention of users. In addition, this study finds new patterns that can potentially help separate health disinformation. For example, we find that less quotation and hyperlinks are more associated with unreliable outlets. To show the effectiveness of the features we discovered, we built supervised classification models using the features which showed satisfactory performance. We also conducted an experiment for measuring the importance of the features which revealed the potentiality of the extracted features in distinguishing reliable sourced health news from unreliable sourced health news. However, there are some limitations to this study. For instance, we didn't consider the videos, cited experts, comments of the users, and other information. In the future, we want to overcome these limitations and leverage the findings of this study to combat health disinformation.

REFERENCES

[1] Chris Breaux. (accessed September 28, 2018). *"You'll Never Guess How Chartbeat's Data Scientists Came Up With the Single Greatest Headline"*. http://blog.chartbeat.com/2015/11/20/youll-never-guess-how-chartbeats-data-scientists-came-up-with-the-single-greatest-headline/

[2] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.

[3] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing health-related misinformation on social media. *Jama* 320, 23 (2018), 2417–2418.

[4] Nicole K Dalmer. 2017. Questioning reliability assessments of health information on social media. *Journal of the Medical Library Association: JMLA* 105, 1 (2017), 61.

[5] Irja Marije de Jong, Frank Kupper, Marlous Arentshorst, and Jacqueline Broerse. 2016. Responsible reporting: neuroimaging news in the age of responsible research and innovation. *Science and engineering ethics* 22, 4 (2016), 1107–1130.

[6] Juliette De Maeyer. 2012. The journalistic hyperlink: Prescriptive discourses about linking in online news. *Journalism Practice* 6, 5-6 (2012), 692–701.

[7] Katie Forster. (accessed October 30, 2018). *Revealed: How dangerous fake health news conquered Facebook*. https://www.independent.co.uk/life-style/health-and-families/health-news/fake-news-health-facebook-cruel-damaging-social-media-mike-adams-natural-health-ranger-conspiracy-a7498201.html

[8] Susannah Fox. (accessed October 30, 2018). *The social life of health information*. http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/

[9] Gaby Galvin. (accessed October 30, 2018). *How Bots Could Hack Your Health*. https://www.usnews.com/news/healthiest-communities/articles/2018-07-24/how-social-media-bots-could-compromise-public-health

[10] Bryan Gardiner. (accessed September 28, 2018). *"You'll Be Outraged at How Easy It Was to Get You to Click on This Headline"*. https://www.wired.com/2015/12/psychology-of-clickbait/

[11] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*. IEEE, 518–518.

[12] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. In *2018 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*. ACM.

[13] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 729–736.

[14] Muiris Houston. (accessed October 31, 2018). *Measles back with a vengeance due to fake health news*. https://www.irishtimes.com/opinion/measles-back-with-a-vengeance-due-to-fake-health-news-1.3401960

[15] informationisbeautiful.net. 2016. Unreliable/Fake News Sites & Sources. https://docs.google.com/spreadsheets/d/1xDDmbr54qzzG8wUrRdxQl$_C$1dixJSIYqQUaXVZBqsJs.

[16] Mathew Ingram. (accessed October 30, 2018). *The internet didn't invent viral content or clickbait journalism âĂŤ there's just more of it now, and it happens faster*. https://gigaom.com/2014/04/01/the-internet-didnt-invent-viral-content-or-clickbait-journalism-theres-just-more-of-it-now-and-it-happens-faster/

[17] American Press Institute and the Associated Press-NORC Center for Public Affairs Research. (accessed September 28, 2018). *The Personal News Cycle: How Americans choose to get their news*. https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/

[18] Marjorie Kagawa-Singer and Shaheen Kassim-Lakha. 2003. A strategy to reduce cross-cultural miscommunication and increase the likelihood of improving health outcomes. *Academic Medicine* 78, 6 (2003), 577–587.

[19] Alexander Kinsora, Kate Barron, Qiaozhu Mei, and VG Vinod Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*. IEEE, 456–461.

[20] Patty Kostkova, Vino Mano, Heidi J Larson, and William S Schulz. 2016. Vac medi+ board: Analysing vaccine rumours in news and social media. In *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 163–164.

[21] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[22] Felisbela Lopes, Teresa Ruão, Zara Pinto Coelho, and Sandra Marinho. 2009. Journalists and health care professionals: what can we do about it?. In *2009 Annual Conference of the International Association for Media and Communication Research (IAMCR),"Human Rights and Communication"*. 1–15.

[23] Amr Magdy and Nayer Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 103–110.

[24] Saurabh Mathur. (accessed September 24, 2018). *Clickbait Detector*. https://github.com/saurabhmathur96/clickbait-detector

[25] HEALTH ON THE NET. (accessed September 24, 2018). . https://www.hon.ch/en/

[26] Matthew Opatrny. (accessed September 28, 2018). *"9 Headline Tips to Help You Connect with Your Target Audience"*. https://www.outbrain.com/blog/9-headline-tips-to-help-marketers-and-publishers-connect-with-their-target-audiences/

[27] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[28] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 232–239.

[29] Md Main Uddin Rony, Mohammad Yousuf, and Naeemul Hassan. 2018. A Large-scale Study of Social Media Sources in News Articles. *arXiv preprint arXiv:1810.13078* (2018).

[30] Michael Schneider. (accessed September 24, 2018). *Most-Watched Television Networks: Ranking 2016's Winners and Losers*. https://www.indiewire.com/2016/12/cnn-fox-news-msnbc-nbc-ratings-2016-winners-losers-1201762864/

[31] Gary Schwitzer. 2008. How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. *PLoS medicine* 5, 5 (2008), e95.

[32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[33] Miriam Shuchman and Michael S Wilkes. 1997. Medical scientists and health news reporting: a case of miscommunication. *Annals of Internal Medicine* 126, 12 (1997), 976–982.

[34] S Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly* 75, 1 (1998), 55–68.

[35] Emily K Vraga and Leticia Bode. 2017. Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication* 39, 5 (2017), 621–645.
[36] World Health Organization (WHO). (accessed September 24, 2018). . http://www.who.int/
[37] Wikipedia. (accessed September 24, 2018). *List of fake news websites.* https://bit.ly/2moBDvA
[38] Wikipedia. (accessed September 24, 2018). *Wikipedia:Zimdars' fake news list.* https://bit.ly/2ziHafj
[39] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7, 7 (2014), 589–600.
[40] Melissa Zimdars. 2016. My 'fake news list' went viral. But made-up stories are only part of the problem. https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem.

A Linked Data Model for Facts, Statements and Beliefs

Ludivine Duroyon
Univ Rennes, Inria, CNRS, IRISA
Lannion, France
ludivine.duroyon@irisa.fr

François Goasdoué
Univ Rennes, Inria, CNRS, IRISA
Lannion, France
fg@irisa.fr

Ioana Manolescu
Inria and LIX (UMR 7161, CNRS and Ecole Polytechnique)
Palaiseau, France
ioana.manolescu@inria.fr

ABSTRACT

A frequent journalistic fact-checking scenario is concerned with the analysis of statements made by individuals, whether in public or in private contexts, and the propagation of information and hearsay ("who said/knew what when"). Inspired by our collaboration with fact-checking journalists from Le Monde, France's leading newspaper, we describe here a Linked Data (RDF) model, endowed with formal foundations and semantics, for describing *facts, statements*, and *beliefs*. Our model combines temporal and belief dimensions to trace propagation of knowledge between agents along time, and can answer a large variety of interesting questions through RDF query evaluation. A preliminary feasibility study of our model incarnated in a corpus of tweets demonstrates its practical interest.

CCS CONCEPTS

• **Information systems → Resource Description Framework (RDF)**; *Temporal data*; • **Computing methodologies → Reasoning about belief and knowledge**.

KEYWORDS

Data journalism, linked data, fact-checking, tweet analysis

ACM Reference Format:
Ludivine Duroyon, François Goasdoué, and Ioana Manolescu. 2019. A Linked Data Model for Facts, Statements and Beliefs. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316737

1 INTRODUCTION

An interesting class of questions investigated in journalistic fact-checking is the analysis of *who said what when*. Such an analysis may be made in order to determine where a public figure up for (re)election stands with respect to a given issue (a famous example is John Kerry's Senate voting history on the war in Irak[1]), or the public positions of members of a whole political party on an issue (e.g., the projected Wall between the US and Mexico[2]). Statements are made by individuals or organizations, on certain topics, and typically claim to refer to (real-world) facts. It is common for different actors to make different statements about the same fact or about each other statements. An actor may also make different statements about the same thing at different points in time. Professional standards of journalistic work lead to a high interest in modeling and being able to show statement *sources*, which extends our (informal) definition of data of interest to: *who said what when where*. The source can be public (e.g., a speech whose transcript is available on the Web, or a tweet) or it can be private (e.g., an email that journalists acquire through their sources, or which they create, e.g., transcribing a conversation with a source).

Many free or commercial systems and many research prototypes allow analyzing online media to answer specific questions, for instance, CrowdTangle allows to monitor social media and extract events, Twitonomy and TwitterAnalytics are specifically devoted to analyzing Twitter content etc. In this work, we establish a *generic data model for real-world facts, statements, and beliefs*, including (but not limited to) those expressed on social networks; after modeling data in this way, an endless spectrum of fine-granularity analyses can be applied by leveraging a database management system. *Time* plays an important role in our setting, since we must capture when events occur (or when facts hold), and when different statements are made about them; this serves, for instance, to track position reversals in time, or to keep track of promises[3]. Therefore, we incorporate and extend temporal database elements into our model. Further, we take inspiration from classical AI techniques for modeling *agents' beliefs* in order to correctly reflect the connections between actors and their statements. Finally, we adopt the W3C's Resource Description Framework (RDF) concrete graph data model to make databases described in our model *easy to share and combine (link)* with any other RDF dataset, e.g., one that classifies actors according to their political opinions, connections to companies etc., to enable even more analyses.

Beyond being "white-box" (as opposed to models not publicly shared, used by existing media analysis tools), the biggest advantage of our proposed model is to be *comprehensive* (modeling all the above aspects: facts, agents, beliefs, and information propagation), *interoperable* (being RDF), *extensible* (other data sources can be turned into instances of our model) and endowed with *formal semantics*. Our second contribution, beside the model, is to show how we can exploit it through *queries*, both interesting and feasible with off-the-shelf tools.

The remainder of this paper is organized as follows. We recall some preliminaries, present our model, and illustrate interesting queries on its instances. We describe preliminary experiments in a concrete use case, briefly discuss related works, then conclude.

[1] https://www.factcheck.org/2013/09/kerry-spins-his-record-on-iraq/
[2] https://www.politifact.com/truth-o-meter/statements/2019/jan/09/donald-trump/trump-democrats-reverse-border-wall-position/
[3] https://www.lemonde.fr/les-decodeurs/visuel/2018/05/07/un-an-apres-son-election-emmanuel-macron-tient-il-ses-promesses-de-campagne_5295281_4355770.html

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316737

RDF statement	Triple	Short notation
Class assertion	$(s, \text{rdf:type}, o)$	(s, τ, o)
Property assertion	(s, p, o) with $p \neq \text{rdf:type}$	(s, p, o)

Table 1: RDF assertions.

2 PRELIMINARIES

We present next the notions of RDF graphs [16] and how they can be queried with the popular conjunctive fragment of SPARQL [17].

RDF graphs. An RDF graph is a set of triples (s, p, o) where s is termed the *subject*, p the *property*, and o the *object*; such a triple states that s is described with the property p that has value o. *Well-formed* triples, as per the RDF specification, belong to $(\mathcal{I} \cup \mathcal{B}) \times \mathcal{I} \times (\mathcal{I} \cup \mathcal{L} \cup \mathcal{B})$, where $\mathcal{I}$ is a set of *Internationalized Resource Identifiers* (IRIs in short), $\mathcal{L}$ is a set of *literals* (constants), and $\mathcal{B}$ is a set of *blank nodes* that, similarly to labeled nulls [1, 9], representing unknown IRI or literal values.

Notation. Blank node names start with the letter b and literals are written as string between quotes, e.g., "string".

An RDF graph models a set of *assertions*, each of which expresses either an instance of *class* (unary relation) or an instance of *property* (binary relation). The syntax of these assertions is shown in Table 1; we introduce the shorthand notation τ to denote the standard rdf:type property, recommended by the W3C for specifying the class(es) to which a resource belongs.

Example 2.1 (RDF triples). The facts comprised in the phrase "Presidential candidate Trump visits Moscow; Michael D. Cohen works for him"[4] can be modeled with the three triples: (D.Trump, candidateFor, president), (D.Trump, visits, Moscow) and (M.D.Cohen, worksFor, D.Trump).

Querying RDF. SPARQL is the W3C's standard to query RDF graphs. We consider the popular SPARQL conjunctive queries, a.k.a. Basic Graph Pattern Queries (or BGPQs, in short), a core subset of SPARQL 1.1. In particular, we consider BGPQs extended with *property paths*.

A BGPQ q is of the form $q(\bar{x}) \leftarrow t_1, \ldots, t_n$ where $t_1, \cdots, t_n$ are generalized triples called *triple patterns*, in which variables may be used as subject, property or object; $\bar{x}$ is the set of q's *answer variables*, which is a subset of the variables used in $t_1, \cdots, t_n$.

Example 2.2 (BGPQ). The query below asks for the presidential candidates and their collaborators:

$$q(x, y) \leftarrow (\text{x}, \text{candidateFor}, \text{president}), (\text{y}, \text{worksFor}, \text{x})$$

Its answer on the triples in Example 2.1 is: (D.Trump, M.D.Cohen).

Further, in our BGPQs, we allow the use of a property paths in the property position of triple patterns. A property path is recursively defined as either a URI, a variable, or a regular expression on property paths among: $p_1|p_2$ for alternative property paths, p_1/p_2 for a sequence of property paths, $p_1?$ for a property path that occur at most once, p^+ for a property path that occur at least once, and p^* for a property path that may not occur or may occur at least once.

[4]This and the next examples are inspired from https://www.nytimes.com/interactive/2019/01/26/us/politics/trump-contacts-russians-wikileaks.html.

Example 2.3 (BGPQ with property path). The query below asks for those who work for sombody visiting Moscow:

$$q(x) \leftarrow (\text{x}, \text{worksFor/visits}, \text{Moscow})$$

Its answer on the triples in Example 2.1 is: M.D.Cohen.

3 DATA MODEL

We now describe our data model capable of modeling (timed) beliefs, facts and statements.

3.1 Agents, Time, Facts and Beliefs

Agents can be individuals, organizations (e.g., companies, media etc.), or other "party" which make statements or learn about them. In the following, we model agents as RDF resources; for simplicity of the examples, we may simply use people names to denote them, e.g. Alice, Bob etc. In general, agents are those resources having the type *Agent*.

Time We consider a set T of *time intervals*, each of which has a *start* time and an *end* time. Further, we distinguish three special constants $-\infty$, *now*, $+\infty$ that are used to model time bounds $(-\infty, +\infty)$, as well as the current time. A finite union of time intervals can always be equivalently written as a *disjoint* union of time intervals. By a slight abuse of notation, in the sequel, we will write $t \in T$ to denote any such finite union of disjoint time intervals. Each start time and end time can be represented e.g. following the W3C's XML Schema `dateTime` type[5], under the form YYYY-MM-DD[THH:MM].

Temporal normalization To simplify working with unions of time intervals, we use their normalized forms, which are equivalent smallest unions of non overlapping intervals [5]:

Definition 3.1 (Time normalization). Let $I = \{\iota_1, \ldots, \iota_k\}$ be a set of temporal intervals. The time normalization of I is the set of intervals $I' = \{\iota'_1, \ldots, \iota'_l\}$, where l the smallest integer such that:

- $\iota'_1, \ldots, \iota'_l$ are pairwise disjoint and
- $\iota_1 \cup \ldots \cup \iota_k = \iota'_1 \cup \ldots \cup \iota'_l$.

It is easy to see that I' is well-defined, that is: for any I, I' exists and it is unique. Clearly, I' has at most as many intervals as I, i.e., $l \leq k$. For instance, if $k = 3$, $\iota_1 = [1998, 1999]$, $\iota_2 = [2002, 2005]$ and $\iota_3 = [2003, now]$, we have $l = 2$, $\iota'_1 = [1998, 1999]$ and $\iota'_2 = [2002, now]$. We call *temporal normalization* the procedure which given a set of intervals $\{\iota_1, \iota_2, \ldots, \iota_k\}$ computes the disjoint set of intervals $\{\iota'_1, \iota'_2, \ldots, \iota'_l\}$ described above. From now on, we consider that time is normalized set of intervals; we use ι, possibly with indices, to denote them.

We model intervals as RDF ressources having the type Interval. Each interval has a *begin* and an *end* property specifying the interval bounds, whose values are either of type `dateTime`, or a special constant called *now*. To represent time as a set of intervals, we use an RDF ressource of type Time and a property *hasInterval* taking values of type Interval.

Example 3.2. The time t_0 is the union of the two intervals ι_0 = [2018-04-07T16:00;2018-04-07T19:30], i.e., that starts on April 7, 2018 at 4:00 PM and ends on April 7, 2018 at 7:30 PM, and ι_1 = [2018-04-07T21:00;2018-04-07T22:30]. This is stated in RDF as: $(\text{t}_0, \tau, \text{Time})$, $(\text{t}_0, \textit{hasInterval}, \iota_1)$, $(\text{t}_0, \textit{hasInterval}, \iota_2)$

[5]See https://www.w3.org/TR/xmlschema11-2/#dateTime

(ι_1, τ, Interval), (i_2, τ, Interval),
(ι_1, *begin*, 2018-04-07T16:00), (ι_1, *end*, 2018-04-07T19:30),
(ι_2, *begin*, 2018-04-07T21:00)(ι_2, *end*, 2018-04-07T22:30).

For simplicity, if a time is a single interval, we directly associate a begin and an end to a time. For instance if t_0 is only ι_1 then we write : (t_0, *begin*, 2018-04-07T16:00), (t_0, *end*, 2018-04-07T19:30).
Fact is used to designate in a generic way any real-life event recorded in the database. In the following, we denote facts by their IDs F_1, F_2 etc. and model each fact as an RDF resource of type *Fact*. Each fact has a *time* property specifying when the fact is supposed to occur. Further information about the fact itself is the value of the property *description*; this can be e.g., a text, or an RDF resource having more properties etc. We assume that time and sign information are always attached to facts, e.g., an interval $[-\infty, +\infty]$ for facts that are always true etc. If the database features (F, time, t_1) and (F, time, t_2) with $t_1 \neq t_2$, we represent this as a single fact F with time $t_1 \cup t_2$.

Example 3.3 (Fact). The fact "Trump conducted business in Moscow on July 2-3, 2016" is represented as:
(F_1, τ, Fact), (F_1, *time*, t_1), (t_1, *begin*, 2016-07-02),
(t_1, *end*, 2016-07-03), (F_1, *description*, u_1), (u_1, *who*, D.Trump),
(u_1, *what*, conductsBusiness), (u_1, *where*, Moscow).

This small RDF graph together with those corresponding to the next two examples are sketched in Figure 1; oval nodes denote URIs, while text nodes correspond to literals. Some information such as edge labels, triples stating the time of t_1, t_2 etc. are ommitted to avoid clutter; URIs representing agents are signaled by a "user" pictogram. The reason why some nodes have a yellow background will be discussed shortly.
Beliefs relate agents with the things they believe; we model them as resources of the dedicated RDF type Belief. In this work, we use "believe" to model any among: *having knowledge (being informed) of*, *thinking* or *believing* something etc. A belief can be a positive belief (the agent believes something) or a negative one (the agent doesn't believe it). A belief is characterized by: (*i*) the agent holding the belief, which is the value of a *from* property whose subject is the belief; (*ii*) the time when the agent holds the belief, represented by a *time* property; (*iii*) the belief subject, which is the value of a *believes* property, can be a fact (either one we consider to hold, or one of which we only know that it is stated or believed by some agent), a communication, or another belief; (*iv*) finally, a *sign* property whose values can be + or −, indicating whether the agent actually believes the subject of the belief, or not. For simplicity, we assume the *sign* property is present only when its value is −; when it is not present, we assume its value is +, i.e., the agent does hold the belief.

Example 3.4 (Beliefs). J. Mueller believes since January 5, 2017, and up to now that D. Trump conducted business Moscow in July 2016. Building on the fact F_1 from Example 3.3, we denote this by:
(B_1, τ, Belief), (B_1, *sign*, +), (B_1, *from*, J.Mueller), (B_1, *time*, t_2),
(t_2, *begin*, 2017-01-05), (t_2, *end*, *now*), (B_1, *believes*, F_1).

3.2 Belief sharing: communications

Information (beliefs) spread through time through communications. Each communication is characterized by: (*i*) an agent who is the transmitter, indicated by its *from* property; (*ii*) optionally, one or more agents who are the receivers, indicated by the *to* property; (*iii*) a subject, which is the value of the *communicates* property, which can be a fact, a belief or another communication; (*iv*) a sign, whose value is given by a *sign* property + or −, indicates whether the agent actually agrees or disagrees with the subject of the communication; (*v*) a time encoded by the *time* property. If the receiver is not specified, we assume it is a *public* communications. These are considered available to anyone, e.g., anyone can have access to the newspaper, TV, Web source where the communication was made. Communications with one or more specific receivers are considered private; only the transmitter and the receiver are assumed aware of this communication.

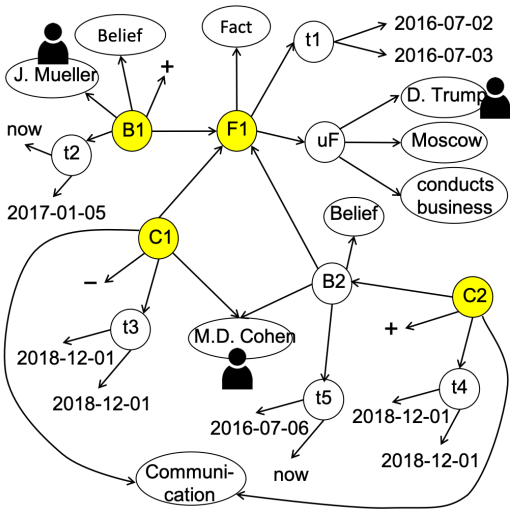


Figure 1: Illustration of the Examples 3.3, 3.4 and 3.5.

Example 3.5 (Communications). On Dec 1st, 2018, M. D. Cohen states that D. Trump did not conduct business in Moscow in July 2016. We denote this by:
(C_1, τ, Communication), (C_1, *sign*, −), (C_1, *time*, t_3),
(t_3, *begin*, 2018-12-01), (t_3, *end*, 2018-12-01),
(C_1, *communicates*, F), (C_1, *from*, M.D.Cohen)
On Dec 15, 2018, J. Mueller says M. D. Cohen knows, since July 2, 2016, that D. Trump conducted business in Moscow in July 2016:
(C_2, τ, Communication), (C_2, *sign*, +), (C_2, *time*, t_4),
(t_4, *begin*, 2018-04-16), (t_4, *end*, 2018-04-16), (C_2, *communicates*, B_2),
(B_2, τ, Belief), (B_2, from, M.D.Cohen), (B_2, believes, F), (B_2, time, t_5),
(t_5, begin, 2016-07-06), (t_5, end, now).

3.3 Records and databases

The above examples show that some facts are considered stated as such in the database, e.g., F_1 in Example 3.3, whereas others are not, hence *not declared true* in the database, but merely *believed* or *communicated by* some agents whether they are actually true or not. For instance, the database may state M. D. Cohen believes that D. Trump was on holidays in Moscow in July 2016, in which case the fact that D. Trump was on holidays in Moscow in July 2016 just holds *according to M. D. Cohen* within the database. Similarly,

the database may state that an agent A holds a belief, or makes a communication, but this is different from the database stating that *according to agent B*, agent A believes, respectively communicates it. In the former case, something holds *according to the database*, i.e., is considered an undisputed assertion; in the latter, the database merely states that something just holds according to B.

Records We introduce a special type Record which we attach to any fact, belief or communication that holds according to the database, i.e., that the database declares to be true. Note that each record may be the root of a potentially long chain of beliefs and communications; each such chain ends in a Fact[6]. We illustrate this below by revisiting the examples above.

Example 3.6 (Records). In Example 3.3, F_1 holds according to the database. Thus, the triple $(F_1, \tau, \text{Record})$ is also part of the encoding of the fact F in our data model; in Figure 1, nodes of type Record are shown on a yellow background.

In Example 3.4, the belief B_1 holds according to the database. Thus, the triples $(B_1, \tau, \text{Record})$ is also part of it.

Similarly, in Example 3.5, the communications C_1, C_2 took place according to the database, hence the triples $(C_1, \tau, \text{Record})$ and $(C_2, \tau, \text{Record})$ are part of it. However, the belief B_2 is not a database record, because it only holds according to J. Mueller, who communicates on it.

Database Based on the elements introduced above, a *database* can be seen as a *set of fact, belief and communication records*, encoded by a set of RDF triples as discussed above. Figure 1 displays such a database, in which F_1, B_1, C_1, C_2 are the database records.

3.4 Use case: political Twitter scenario

As a particular instantiation of this data model, we have built a database for a French political tweet analysis scenario, as follows. We have subscribed to, and archived, tweets from elected officials, ministries, politicians etc.; over Sept-Dec 2018, we obtained a total of 900.000 tweets (**4.57GB** in JSON format). Out of these, we produced an instance of our RDF data model of size **4.5GB**, as follows. Each tweet is a public Communication, by an agent who is the twitter account. Since each tweet has certainly occurred, it is also a Record. The time of the tweet is a point interval (with the beginning equal to its end). What the tweet communicates depends on the nature of the tweet. (*i*) The content of a *simple tweet* is a text message and possibly photos, links,hashtags and/or agents mentioned; all these can be easily extracted from the corresponding fields within the JSON tweet format. We represent such a tweet content by a Fact which is also a Record. Each such Fact has an ID described by the above mentioned features extracted from the JSON tweet, using the following properties: *text* for the tweet text message, *urlPhoto* for the photos, *urlMedia* for the links and *hashtag* for the hashtags. Further, if the tweet t, by agent $@a$, whose content is modeled by the fact f, mentions another agent (Twitter user) $@b$, we create a communication from $@a$ to $@b$, with the same time moment as t, whose content is the same fact f. (*ii*) The other tweets are either *retweets*, *answers* to a tweet, or tweets that *quote* another tweet; each of this is a tweet t_1 based on another tweet t_0. From t_1, we build a Communication with time and author as above; however, it will be based on the communication corresponding to the content of t_0. Further, answers and quotes (but not retweets) may also have some content of their own; we model that by a new Comment subclass of Fact, and an *addsComment* property which goes from the Communication which models t_1, to the actual text of the answer (or quote) t_1.

Example 3.7 (Tweets). Tweet t_0 from the PublicSÃĭnat TV station reads: "#Benalla affair : @CCastaner accuses senators of threatening the republic" with a link to his webpage. A communication c_0 from @publicsenat is created, associated to the fact f_0 of the text, the link, and the hashtag. Another communication c_{0bis} from @publicsenat to @CCastaner on f_0 is created. Twitter user @PadrePio quotes t_0 in his tweet t_1, adding "The new Minister of the Interior, also!" This leads to a communication c_1 from @PadrePio, on c_0 (the declaration of @publicsenat). c_1 has also a comment co_1 with his text.

Further, we used an open dataset about the French National Assembly[7], which we also converted in RDF; this includes the Twitter account information, enabling to connect the tweets to more information about their authors. We added a database of French politician Twitter accounts that journalists from Le Monde shared with us, leading to a total of 48.877 triples describing French political Twitter users.

Our complete corpus (dominated by the triples describing tweets) has 20.697.338 RDF triples. All its Facts and Communications are Records, i.e., there is no "hearsay" (nothing is hypothetical). However, this scenario is sufficient to enable studying the performance of queries that go along chains of Communications: these are materialized by retweets, answers, and/or quotes.

4 QUERYING OUR DATA MODEL

Many interesting queries can be written on an instance of our data model. A few examples include: (*i*) What did M. X. state on date D in his interview with journal J? (*ii*) What is the fact on which members of the political party PA communicate most often, that members of the political party PB do not mention at all? (*iii*) What are the facts on which M. X and Mrs. Y disagree (he believes it whereas she does not)? (*iv*) Which are the facts from a given knowledge base KB that Mrs. Y does not believe (belief with a − sign)? (*v*) Who changed her/his mind, or made incoherent statements, on some fact within a time period (beliefs, or communications, with opposite signs but on a same fact)? (*vi*) What are the informations (beliefs, communications) shared with M. X about fact F before the time moment T? Such queries can be written directly by someone who knows SPARQL, or formulated with the help of a GUI which allows specifying how many facts (beliefs, communications) the query is about, how they connect, fill in the known values (e.g., X, D, J, Y etc. in the above examples), then the GUI generates the corresponding SPARQL statement.

Below, we highlight a family of queries that are interested in tracing the *propagation* of information across chains of beliefs and communications. These queries are theoretically interesting as they translate to property paths, a relatively relatively recent SPARQL

[6]This follows the natural interpretation that any belief or communication carries over something, thus the chain must end in a Fact.

[7]https://github.com/regardscitoyens/nosdeputes.fr/blob/master/doc/api.md#liste-des-parlementaires

feature, yet, as we will show, they are feasible for RDF engines freely available today.

We start with a set of **generic queries**, which can be written on any instance of the data model previously described. Q_0 captures who has heard of what, when, and how. We consider an agent *hears about* something when either the agents believes it, or the agent is the recipient of a communication about it. Q_0 returns: the agent a, the communication c, its time (as an interval spanning from b to e), and its subject s. Note the regular path expression (of potentially unbounded length) going from c to the subject s: it captures the propagation of hearsay, i.e., if c is about a belief b (or another communication c') which is about topic t, then Q_0 returns t together with a and c, since it is through c that a heard of t:

$$\begin{aligned} \mathtt{Q}_0(?a, ?c, ?b, ?e, ?s) \leftarrow & (?c, \tau, \text{Record}), \\ & (?c, (from|to), ?a), \\ & (?c, (believes|communicates)+, ?s) \\ & (?c, time, ?t), (?t, begin, ?b), (?t, end, ?e) \end{aligned}$$

Building on Q_0, Q_1 finds out who has heard of a specific fact f, and when. It returns the agents, the communications which propagated f to them, and the communication time:

$$\mathtt{Q}_1(?a, ?c, ?b, ?e) \leftarrow \mathtt{Q}_0(?a, ?c, ?b, ?e, f), (f, \tau, \text{Fact})$$

Q_2 is a selection on Q_0. It finds what a given agent a has heard of, how, and when:

$$\mathtt{Q}_2(?c, ?b, ?e, ?s) \leftarrow \mathtt{Q}_0(a, ?c, ?b, ?e, ?s)$$

Q_3 finds who has heard of what, through which communication, during a given time interval $t0$=$[b0, e0]$:

$$\begin{aligned} \mathtt{Q}_3(?a, ?c, ?s) \leftarrow \quad & \mathtt{Q}_0(?a, ?c, ?b, ?e, ?s), OVERLAP([b, e], [b0, e0]) \\ & FILTER(?b2 > b), FILTER(?e2 < e) \end{aligned}$$

Next, we present a few queries **specific to our political Twitter scenario**. To find out how a given hashtag h disseminated on Twitter, query Q_4 returns all the agents reached by the hashtag and the communication which brought the hashtag to them:

$$\mathtt{Q}_4(?a, ?c, ?b, ?s) \leftarrow \mathtt{Q}_0(?a, ?c, ?b, ?s), (?s, hashtag, h)$$

Because a tweet has a zero-length temporal interval (b=e), in Q_4 we only use the interval beginning. A restriction of Q_4, query Q_5 ensures that c has taken place in a given time interval:

$$\begin{aligned} \mathtt{Q}_5(?a, ?c, ?f) \leftarrow & \mathtt{Q}_0(?a, ?c, ?b, ?s), (?s, hashtag, h) \\ & FILTER(?b2 > b), FILTER(?b2 < b) \end{aligned}$$

Finally, query Q_6 uses the political party affiliations of agents to aggregate the hashtag targets (agents) by their political party:

$$\begin{aligned} \mathtt{Q}_6(?pa, count(*)\ as\ ?n) \leftarrow & \mathtt{Q}_0(?a, ?c, ?b, ?s), (?s, hashtag, h) \\ & FILTER(?b2 > b), FILTER(?b2 < b) \\ & (?a, name, ?name), (?a, party, ?pa), \\ & GROUP\ BY\ ?pa \end{aligned}$$

5 EXPERIMENTS

We describe experiments we carried out with instances of our data model built as we explained in Section 3.4.

Platform and settings We aimed to check the feasibility of loading and querying instances of our data model using an off-the-shelf RDF data management system (RDB, in short). We had a particular interest in the capacity of the RDB to handle property path queries. Following a recent benchmark of property path support [13], we have chosen **RDF4J** (formerly known as Sesame) v2.4.2. RDF4J provides B+-tree indexes over the stored triple; an index is defined by an order over four attributes s, p, o (for the standard RDF subject, property, object) and c (for context, i.e., the RDF graph from which a triple comes). By default, the *spoc* and *posc* indexes are built; the application can modify the index set. We added three more indexes *psoc*, *cosp*, and *opsc* to speed up query evaluation. We ran our tests on a Linux computer with an Intel Xeon CPU @ 2.67GHz, 4 Cores and 40GB RAM.

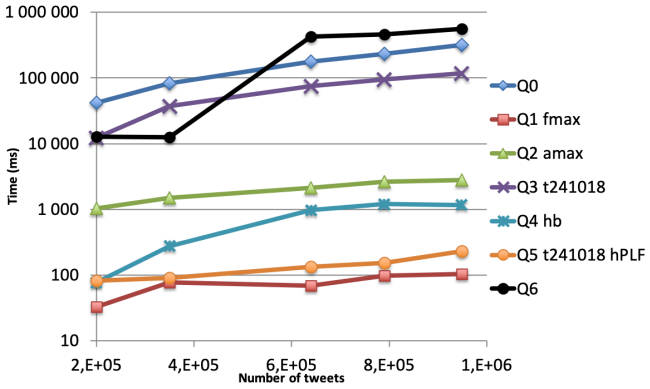


Figure 2: Query time for increasing numbers of tweets.

Loading We loaded our corpus of 2×10^7 triples and built the six indexes on it (two default and four we added) in 976 seconds.

Queries We timed the execution of a set of queries on increasingly larger subsets of our dataset, from 1/6th of the it (150.000 tweets) to the entire dataset obtained from 900.000 tweets; the query evaluation times (in ms) appear in Figure 2.

As a first baseline, we ran Q_0 (from Section 4); this query traverses all the chains of information propagation in our dataset, and returns a very large answer (from 1 million to tens of millions of answers); it is not very interesting per se, but it is one of the most expensive we could think of (taking 318 seconds when run over the whole dataset), yet RDF4J is capable of handling it. A close inspection of the times shows that Q_0 scale-up is *super-linear in the number of tweets (database size)*, but it is *linear in the number of results*, which is overall good news (the query engine can hardly escape the cost of enumerating results).

Next, we ran instantiations of the query templates introduced in the previous section, as follows. First, we manually identified the fact $f_{\max}$ which is at the origin of the longest chain of Communications (retweets, comments etc.). Then, we ran Q_1 setting f to $f_{\max}$: find everyone who heard of this fact directly or indirectly; this query has 374 results. Thanks to our indexes, Q_1 with f=$f_{\max}$ ran extremely fast (from 50 ms to 99 ms, the lowest curve in the figure). This is very good news, as it shows that RDF4J is capable of restricting the exploration of "has heard of" chains to just those ending in $f_{\max}$; the query is three orders of magnitude faster than Q_0. However, using just the default indexes, Q_1 performed no better than Q_0! This is because with our modeling (recall Q_0 from Section 4, and also Figure 1), information propagation chains go "backward" (a fact is the object of a triple connecting it to a belief or communication about it, which in turn are objects (property values) of higher-order beliefs or communications etc). This is why the indexes we added to RDF4J made an important difference here.

Subsequently, we picked an agent $a_{\max}$ who authored the most tweets in the smallest (scale 1/6) fragment of our database, then ran

Q_2 setting a to $a_{\max}$. Q_2 takes more time than Q_1, since it needs to look for that agent at many levels in the information propagation chains; however, it is much faster than Q_0, (also) because it returns less results. Next, we ran Q_3 for the time interval t_{241018} from 9 to 10 am on Oct 24, 2018; it is quite expensive, showing that the time selection was not very efficiently exploited by RDF4J to prune the search. The tweet dates are triple objects, thus we hoped that indexes starting with the o attribute would be exploited by RDF4J to speed up significantly these queries, or, this does not appear to be the case. RDF4J does not provide an "explain" functionality, thus for now we can only guess at an imperfect optimization strategy.

We picked the hashtag #Benalla (a name involved in a political scandal in France), which we denote hb in the sequel, and ran Q_4 setting h to hb; Q_4 is moderately expensive, as it also had to traverse all "has heard of" chains starting in a relevant Communication We ran Q_5 using t_{241018} and a popular hashtag h_{PLF} of the moment, PLF19[8]; combining the selections on the hashtag and on the time significantly reduced the amount of data manipulated by the engine, thus Q_5 performed almost as well as Q_1.

Finally, we ran Q_6 with h set to #PLF2019 and the time interval t_{241018}. This query becomes more expensive than Q_0 for large datasets; it combines the disadvantage of (not-so-efficient) temporal selection and the extra cost of joining the communications with the agent political party affilations, and of computing the group-by.

From the experiments, we conclude that RDF4J handles reasonably well complex queries, featuring complex path expressions (which lead to exploring potentially long information dissemination chains). Our study focused on such queries since information propagation is both interesting for data journalism scenarios and a recent, challenging SPARQL feature. While we used publicly shared Tweets, such queries can be great tools in investigative journalism scenarios based (also) on information accessible only to journalists. Modeling data in RDF aims (also) at facilitating the integration of all the data sources journalists can acquire.

6 RELATED WORK AND CONCLUSION

Modelling facts, statements and beliefs to further search and analyze them has recently gained interest in the setting of computational journalism [4], especially for fact-checking purposes, like monitoring sources, extracting claims, checking them w.r.t. reference sources and publishing obtained results (see [2] for a survey).

In this setting, we follow a *database-oriented approach* for representing timed facts, statements and beliefs using the RDF data model, and then exploiting them through SPARQL queries.

Belief databases [6], representing (*i*) facts and (*ii*) positive or negative beliefs of agents on facts or on agent beliefs, have been investigated in the setting of the relational data model and conjunctive queries. Such databases build on multi-agent epistemic logic, in particular to infer and query (a possibly infinite set of) implicit agent beliefs. We do not consider belief inference; we focus on storing information and how it propagates between agents.

Temporal databases, long known for relational data [14], have been revisited for RDF. [10] attach time points or time intervals to triples; [12] allows intervals with unknown bounds on which constraints can be set or derived using Allen's algebra. Finally, [15] models a timed RDF graph as a set of its snapshots over time. We build on the time modelling approach of [12] to attach a normalized union of time intervals to facts, beliefs and communications.

RDF has also been used in fact-checking. In [3], claims are RDF triples to be checked against a knowledge base such as DBPedia. Claim accuracy is assessed based on the (shortest and most specific) paths that connect the claim subject to its object within the knowledge base. FactMinder [7] uses an XML-RDF hybrid data model [8] to link the entities found in documents (web pages, etc.), hence in the claims made there, to these entity descriptions in a knowledge graph, in order to guide manual fact-checks. Finally, [11] proposes a complete fact-checking systems, from claim extraction to analysis and publication of the results, which notably check claims against several knowledge bases. However, it does not use a fine-grained temporal model for facts, beliefs, and their propagation.

As the next step of this work, we are currently devising a form-based GUI to help journalists writing meaningful but complex queries for their investigations.

[8] *Projet de la loi de finance* 2019, or the French state budget for 2019.

ACKNOWLEDGMENTS

This work is funded by the ANR ContentCheck project (contract ANR-15-CE23-0025).

REFERENCES

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. 1995. *Foundations of Databases*. Addison-Wesley.
[2] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A Content Management Perspective on Fact-Checking. In *WWW*. https://doi.org/10.1145/3184558.3188727
[3] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015).
[4] Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational Journalism. *Commun. ACM* 54, 10 (2011). https://doi.org/10.1145/2001269.2001288
[5] Andreas Dohr, Christiane Engels, and Andreas Behrend. 2018. Algebraic Operators for Processing Sets of Temporal Intervals in Relational Databases. In *TIME*. https://doi.org/10.4230/LIPIcs.TIME.2018.11
[6] Wolfgang Gatterbauer, Magdalena Balazinska, Nodira Khoussainova, and Dan Suciu. 2009. Believe It or Not: Adding Belief Annotations to Databases. *PVLDB* 2, 1 (2009). http://www.vldb.org/pvldb/2/vldb09-115.pdf
[7] François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. 2013. Fact checking and analyzing the web. In *SIGMOD*. https://doi.org/10.1145/2463676.2463692
[8] François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. 2013. Growing triples on trees: an XML-RDF hybrid model for annotated documents. *VLDB J.* 22, 5 (2013). https://doi.org/10.1007/s00778-013-0321-2
[9] François Goasdoué, Ioana Manolescu, and Alexandra Roatis. 2013. Efficient query answering against dynamic RDF databases. In *EDBT*.
[10] Claudio Gutiérrez, Carlos A. Hurtado, and Alejandro A. Vaisman. 2005. Temporal RDF. In *ESWC*. https://doi.org/10.1007/11431053_7
[11] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. *Proceedings of the VLDB Endowment* 10, 7 (2017).
[12] Carlos A. Hurtado and Alejandro A. Vaisman. 2006. Reasoning with Temporal Constraints in RDF. In *Principles and Practice of Semantic Web Reasoning (PPSWR)*. https://doi.org/10.1007/11853107_12
[13] Daniel Janke, Adrian Skubella, and Steffen Staab. 2017. Evaluating SPARQL 1.1 Property Path Support. In *2nd Int'l Workshop on Benchmarking Linked Data, next to ISWC*. http://ceur-ws.org/Vol-1932/paper-04.pdf
[14] Christian S. Jensen and Richard T. Snodgrass. 1999. Temporal Data Management. *IEEE Trans. Knowl. Data Eng.* 11, 1 (1999). https://doi.org/10.1109/69.755613
[15] Jonas Tappolet and Abraham Bernstein. 2009. Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL. In *ESWC*. https://doi.org/10.1007/978-3-642-02121-3_25
[16] w3c-rdf [n. d.]. RDF. https://www.w3.org/TR/rdf11-concepts.
[17] w3c-sparql [n. d.]. SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query.

Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking

Casper Hansen
Department of Computer Science,
University of Copenhagen

Christian Hansen
Department of Computer Science,
University of Copenhagen

Stephen Alstrup
Department of Computer Science,
University of Copenhagen

Jakob Grue Simonsen
Department of Computer Science,
University of Copenhagen

Christina Lioma
Department of Computer Science,
University of Copenhagen

ABSTRACT

Automatic fact-checking systems detect misinformation, such as fake news, by (i) selecting *check-worthy* sentences for fact-checking, (ii) gathering related information to the sentences, and (iii) inferring the factuality of the sentences. Most prior research on (i) uses hand-crafted features to select check-worthy sentences, and does not explicitly account for the recent finding that the top weighted terms in both check-worthy and non-check-worthy sentences are actually overlapping [15]. Motivated by this, we present a neural check-worthiness sentence ranking model that represents each word in a sentence by *both* its embedding (aiming to capture its semantics) and its syntactic dependencies (aiming to capture its role in modifying the semantics of other terms in the sentence). Our model is an end-to-end trainable neural network for check-worthiness ranking, which is trained on large amounts of unlabelled data through weak supervision. Thorough experimental evaluation against state of the art baselines, with and without weak supervision, shows our model to be superior at all times (+13% in MAP and +28% at various Precision cut-offs from the best baseline with statistical significance). Empirical analysis of the use of weak supervision, word embedding pretraining on domain-specific data, and the use of syntactic dependencies of our model reveals that check-worthy sentences contain notably more identical syntactic dependencies than non-check-worthy sentences.

KEYWORDS

Fact checking; Check worthiness; Deep learning; Weak supervision.

ACM Reference Format:
Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316736

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316736

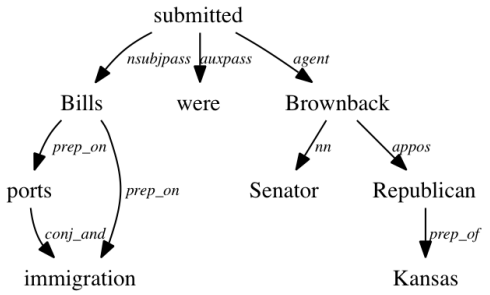


Figure 1: Syntactic dependencies example (from [20]).

1 INTRODUCTION

The fast and wide spread of misinformation (as opposed to true information) on social media [22, 25], and the increasing use of social media as a source of news[1] has turned "fake news" into an important societal problem on a scale that requires automated solutions. An automated fact-checking [21] pipeline typically consists of three steps: (i) selecting *check-worthy* sentences (i.e. sentences that contain check-worthy claims and should be fact-checked), (ii) gathering related information to those sentences, and (iii) using this related information to infer the factuality of each check-worthy sentence. Prior research has so far focused mainly on steps (ii) and (iii), for instance by generating claim-specific queries and querying search engines for relevant supporting information [12], focusing on specific sources such as Twitter [1], or exploiting knowledge graphs from e.g. Wikipedia [5]. These approaches assume an input of check-worthy claims. Considerably less research has focused on detecting such check-worthy claims, that is, determining not whether a sentence is true or not, but whether a sentence contains a check-worthy claim (and should be fact-checked) or not.

Most research on check-worthiness detection uses hand-crafted features, such as bag-of-word representations, sentiment, and embedding averages [7, 8, 10, 19]. In addition, most work in this area does not explicitly account for the recent finding that the top weighted terms in both check-worthy and non-check-worthy sentences are actually overlapping [15], hence compromising the effectiveness of bag-of-word based methods.

Motivated by the above, we present a neural check-worthiness sentence ranking model that uses a dual sentence representation: each word in a sentence is represented by *both* its embedding (aiming to capture the semantics of that word from its context) and its

[1] https://www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from-social-media-survey-idUSKCN1BJ2A8

syntactic dependencies (aiming to capture the role of that word in modifying the semantics of other words in the sentence, see Figure 1). We train the network with these dual representations end-to-end. This allows to learn such descriptive features directly from the input data, rather than relying on predetermined hand-crafted features that may not be useful for the task, and hence to adapt the representations to the domain. To tackle the problem of having very little available training data, we use an existing check-worthiness system to weakly label sentences, and we use this weakly labelled dataset to pretrain our neural network. This is inspired by recent strong results [18, 23] in various information retrieval tasks with few labelled data, but large amounts of unlabelled data.

Thorough experimental evaluation against all state of the art baselines on political speeches from the 2016 U.S. election, shows our model to be superior in all comparisons (+13% in MAP and up to +28% at various Precision cut-offs from the best baseline, with statistical significance). We empirically trace this superior performance to the use of syntactic dependencies in the sentence representation, where we find check-worthy sentences to contain notably more identical syntactic dependencies than non-check-worthy sentences. Further analysis shows that the performance benefits of weak supervision increase with the amount of data used, and that embeddings trained on smaller domain-specific data, as opposed to general purpose embeddings trained on the larger Google News corpus, increase effectiveness. In addition, despite using deep learning (a family of models that is generally considered of low interpretability [24]), the attention weighting used by our model on a word level provides humanly interpretable output, where the parts of the sentence that are important for the check-worthiness prediction can be determined.

We **contribute** a competitive and interpretable end-to-end trainable neural network model for check-worthiness ranking, which uses a dual input representation of both word embeddings and syntactic dependencies. Weak supervision is used to pretrain the network on large amounts of unlabelled data.

2 RELATED WORK

Given a sentence (also referred to as statement) as input, ClaimBuster [8, 9] outputs its check-worthiness score by extracting a set of features (sentiment, statement length, Part-of-Speech (POS) tags, named entities, and tf-idf weighted bag-of-words), and training a SVM classifier on these features to predict check-worthiness. Patwari et al. [19] present a system called TATHYA that is based on similar features, but that also includes as contextual features sentences immediately preceding and succeeding the one being assessed, as well as certain hand-crafted POS patterns. Gencheva et al. [7] also extend the feature set used by ClaimBuster to include more contextual features, such as the sentence's position in the debate text, and whether the debate opponent is mentioned. The work by Gencheva et al. has been extended to both English and Arabic [10]. In the recent CLEF 2018 competition on check-worthiness detection [17], Zou et al. [26] came first by using a large set of features (similarly to the above mentioned work), and doing feature selection with both a χ^2-test and a linear SVM using a L1 regularizer.

Prior work on neural networks for check-worthiness has been done by Konstantinovskiy et al. [14], who use InferSent [6] to derive a universal neural sentence representation and then train a logistic regression classifier on top of that. Similarly to our model, this approach also uses neural sentence embeddings. However, unlike our model, this approach uses *universal* sentence representations, whereas we train our model *end-to-end* to learn the representations directly from the input data, making the learning domain-specific.

In the related domain of sentence factuality detection Jimenez and Li [11] present a neural approach with multiple word embeddings. They artificially generate additional non-factual sentences to be used for training to increase robustness. Similarly to ours, this work also presents neural approaches that combine multiple word representations in order to improve performance. However, whereas the infusion of artificially generated non-factual sentences by Jimenez and Li [11] allows weak supervision of a single class, we obtain weak labels independently of the type of sentence (i.e. not on a single class).

3 NEURAL CHECK-WORTHINESS SENTENCE RANKING

Given a set of sentences as input, the aim is to rank them in descending order of check-worthiness. In order to better differentiate between sentences of varying degree of check-worthiness, We cast this as a ranking problem, as opposed to assigning a binary output to each sentence, following prior work [7, 8, 10] . Note that any ranked output can be made binary using an appropriate threshold, in case a subsequent fact-checking pipeline requires binary labelled sentences.

Given a set of sentences to be ranked, our model learns an end-to-end trained representation of each sentence. We describe first the representation of each word in the sentence (Section 3.1), followed by the neural network architecture (Section 3.2).

3.1 Neural sentence representation

Our model uses a dual sentence representation: each word in a sentence is represented by *both* its embedding and by its syntactic dependencies. The word embedding aims to capture the semantics of that word from its context. Embeddings of this type are generally well understood and have been found effective for check-worthiness detection [14]. The syntactic dependencies of a word aim to capture the role of that word in modifying the semantics of other words in the sentence, for instance by being the subject or predicate of the sentence. We use a syntactic dependency parser [4] to map each word to its dependency (as a tag) in relation to the sentence structure, which we then represent as a one-hot-encoding. Dependency parsing is fast using state of the art tools (approximately 14,000 words per second) [4].

Our motivation is that syntactic dependencies may be important for discriminating between common overlapping top-weighted words in both check-worthy and non-check-worthy sentences. The existence of common overlapping top weighted words in check-worthy and non-check-worthy sentences is reported by Le et al. [15] (see Figure 2 of [15] for examples), and to our knowledge, is not explicitly addressed by any prior check-worthiness approach. We posit that while these common top weighted words may not be distinguishable by their word representation, they may be distinguishable by their syntactic role in the sentence.

Table 1: Statistics of the embedding training, evaluation, and weakly labelled datasets.The evaluation dataset uses binary labels, and the weakly labelled dataset continuous labels in the interval $[0, 1]$.

Dataset	#docs	#sents.	sents. length	unique words	mean label
Embed. tr.	15,059	609,322	16.66	86,244	-
Evaluation	7	2,602	14.04	3,694	0.05
Weakly lab.	161	37,732	16.53	13,314	0.24

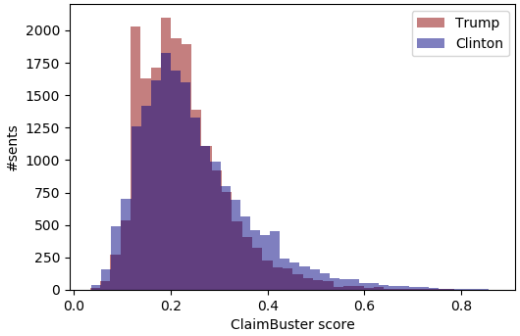


Figure 2: Histogram of the ClaimBuster scores used as the weak labels for each presidential candidate.

3.2 Network architecture

Based on the above, each word in a sentence has two distinct encodings, that together represent the word. We use this double representation of each word as input to a recurrent neural network (RNN) with GRU [3] memory units. The output from each word in the RNN is aggregated using an attention mechanism computed as $\alpha_t = \frac{\exp(\text{score}(h_t))}{\sum_i \exp(\text{score}(h_i))}$, where h_t is the output of the GRU memory cell at time t, and $score(\cdot)$ is a learned function that maps the output to a scalar. The attention-weighted sum is fed to a fully connected layer, from which the output is predicted using a sigmoid activation function. We train the network using the RMSprop optimizer with binary cross entropy as the loss function (details in Section 4.3).

4 EXPERIMENTAL SETUP

We conduct two experiments: (I) we compare our model against state of the art baselines without weak supervision; (II) we use the ClaimBuster API (one of the baselines in experiment I) to weakly label a dataset of unlabelled political speeches (as described in Section 4.2) and we use this weakly labelled data to pretrain the baselines and our model. ClaimBuster API is trained on a non-publicly available dataset and should therefore be considered as a black box baseline.

4.1 Baselines

We compare our model against these baselines (introduced in Section 2), which have yielded state of the art performance at their date of publication: (1) ClaimBuster and its pretrained ClaimBuster API [8], (2) TATHYA [19], and the approaches by (3) Zou et al. [26], (4) Gencheva et al. [7], and (5) Konstantinovskiy et al. [14].

4.2 Data

We use three datasets for three different purposes: (1) the *embedding training* dataset, used to train domain-specific embeddings for our model[2]; (2) the *evaluation* dataset, used to compare our model to the baselines without weak supervision; and (3) the *weakly labelled* dataset, used to compare our model to the baselines with weak supervision. We describe these next (see Table 1 for statistics).

The **Embedding Training Dataset** contains documents related to *all* U.S. elections available through the American Presidency Project[3], e.g. press releases, statements, speeches, and public fundraisers, resulting in 15,059 documents. We use this dataset to pretrain a domain specific word embedding for our model (see Section 5.2).

The **Evaluation Dataset** consists of a total of 2,602 sentences from 7 check-worthiness annotated political speeches from the 2016 U.S election. Out of these 7 speeches, 4 are by Donald Trump and are made available by the CLEF 2018 lab on automatic identification and verification of political claims [17]. The remaining are the inauguration and acceptance speech of Donald Trump and the acceptance speech of Hilary Clinton, and are made available by the authors of ClaimRank [10]. We choose the available PolitiFact annotated labels for this dataset.

The **Weakly Labelled Dataset** consists of all publicly available speeches by Hillary Clinton and Donald Trump from the 2016 U.S. election, which are available through the American Presidency Project. This amounts to 37,732 sentences from 161 speeches not occurring in the evaluation dataset. We use the public API[4] of ClaimBuster [8] to weakly label each sentence in all speeches. The ClaimBuster scores range from 0 to 1 (the higher the score, the more check-worthy the sentence), and are thus continuous as opposed to the binary labels from the evaluation dataset. The distribution of ClaimBuster scores can be seen in Figure 2, where we see that sentences by Hillary Clinton are overall labelled as slightly more check-worthy than those by Donald Trump.

4.3 Tuning

We measure the effectiveness of the ranking outputted by our model and the baselines using mean average precision (MAP), and average precision at multiple cut-offs: P@5, P@10, P@20, and P@R, where R is the number of check-worthy sentences in a given test set. We optimize the MAP when tuning parameters.

We tune and evaluate the approaches using 7-fold cross validation, where the sentences from one speech (see Section 4.2) act as testing data once, while sentences from the remaining speeches are used for training and validation. We use the sentences of each speech as input to the models. In all folds, we set aside 10% of the training data for validation. Each fold-wise evaluation is repeated 5 times with randomly chosen validation data. We report the average score of each metric across the folds and repetitions.

For ClaimBuster [9] and the model by Gencheva et al. [7], we use the best performing setup described in [7]: a double layered fully connected neural network with layer sizes of 200 and 50 respectively. We validate these sizes by keeping the same ratio (4:1) between the layers and test the largest sizes of $\{50, 100, 200, 400\}$,

[2]None of the other approaches support training embeddings.
[3]https://web.archive.org/web/20170606011755/http://www.presidency.ucsb.edu/
[4]https://idir-server2.uta.edu/claimbuster/

test batch sizes of {64, 128, 256, 512}, and use a learning rate of 0.0001. For the approach by Zou et al. [26] we use their multi-layer perceptron model with two hidden layers with sizes of 100 and 8 as per [26]. We validate these sizes by keeping the same ratio (12.5:1) between layers and test the largest sizes of {50, 100, 200, 400} as done earlier. For TATHYA [19] we use the same multi-classifier approach and the same parameters as described in the original paper. For Konstantinovskiy et al. [14] we use the same logistic regression approach as described in the original paper.

For our model, we evaluate the same layer sizes as above with a ratio of 4:1 between the number of neurons in the GRU cell and the single fully connected layer. We train the word embeddings using the word2vec skip-gram model [16] on the embedding training dataset of 15,059 domain specific documents described in Section 4.2. We use standard parameters with a window size of 5 and sample 25 negative samples per word. For the syntactic dependencies of each word, we use the spaCy syntactic dependency parser [4][5].

For the experiment with weak supervision, the ClaimBuster API returns a score from 0 to 1, indicating the degree of check-worthiness estimated by the system. In each fold in the 7-fold cross validation we find the threshold τ that splits the data and makes the fraction of check-worthy samples equal across the training without and with weakly labelled training data. Using these splits we evaluate two thresholding approaches: (1) Binarizing the labels based on τ, and (2) truncating all values larger than τ to the value of τ, and scaling the range $[0, \tau]$ into $[0, 1]$. In the cross validation, our approach performs best with step (2) and the baselines perform best with step (1) (these are the settings we report in Section 5). Note that step (2) can be considered as soft thresholding, as the labels are not binary. The weakly labelled data is used for pretraining the neural models or added to the training data for traditional supervised models.

5 RESULTS

Table 2 shows the results of the experimental comparison of our model to the baselines without and with weak supervision. We see that our model outperforms all baselines across all metrics (with improvements ranging from +9-28%), except P@5 (only without weak supervision) where our model is the second best performing approach. Note that P@k is known to be unstable, especially at small values of k [2, 13]. The best performing baseline is the approach of Konstantinovskiy et al. [14], the only other approach using neural embeddings. This points out the effectiveness of neural word embeddings for this task.

The difference in performance between ClaimBuster and the ClaimBuster API is due solely to the quality of the training data (it is otherwise the same approach) and illustrates the effect of training data quality upon model performance. The fact that our model still notably outperforms the ClaimBuster API baseline shows the benefit of an end-to-end learned representation as opposed to a feature engineered one, for this task.

Only ClaimBuster, the approach of Zou et al. [26], and our model obtain notable improvements from the weakly labelled data (ClaimBuster yields a performance similar to that of the ClaimBuster API). TATHYA [19], and the approaches by Gencheva et al. [7] and Konstantinovskiy et al. [14] do not benefit from weak supervision, most likely because feature-engineering is used as opposed to learning the representation.

5.1 Syntactic dependency similarity between check-worthy sentences

Our syntactic dependencies representations aim to discriminate between top weighted words that are common in check-worthy and non-check-worthy sentences based on the syntactic roles of these words (see Section 3.1). To verify this we compute the average overlap of unique syntactic dependency tags between the following three types of randomly sampled sentence pairs: 1) Pairs of n sampled check-worthy sentences, 2) pairs of n sampled non-check-worthy sentences, and 3) mixed pairs of n sampled check-worthy and n sampled non-check-worthy-sentences. We set $n = 10$ and repeat the computations 1000 times. Table 3 displays the resulting average overlaps and their standard deviation. We observe that check-worthy sentence pairs have the highest average overlap of syntactic dependencies, and non-check-worthy the lowest, but both have a similar standard deviation. This indicates that syntactic dependencies are more similar in check-worthy sentences than in non-check-worthy sentences, and as such constitute a good discriminator between check-worthy and non-check-worthy sentences that otherwise contain an overlap of common top-weighted terms. Note that the average overlap of 7 common syntactic dependencies in check-worthy sentences practically applies to approximately half of the words in those sentences (the average sentence length is 14.04 in that dataset – see Table 1). Mixed sentences (both check-worthy and non-check worthy) have an average overlap in between that of check-worthy and non-check-worthy sentences, but with a larger standard deviation, indicating that syntactic dependencies from this mixed group act as a mixed and less stable discriminating signal.

As an example of the overlap problem of common top-weighted terms, consider the check-worthy sentence "*Since president Obama came into office another two million hispanic americans have fallen into poverty*" and non-check-worthy sentence "*I'm running to be a president for all americans*". In these cases the words *president* and *americans* are both important to describe the content, but have different syntactic dependencies (`compound`/`attr` and `nsubj`/`pobj`, respectively).

5.2 Impact of pretrained word embeddings

Our model is the only approach in Table 2 to have word embeddings trained on a domain specific dataset. All other approaches either use no word embeddings (ClaimBuster [8] and TATHYA [19]), use global word embedding averages (Zou et al. [26] and Gencheva et al. [7]), or use a universally trained sentence representation based on global embeddings (Konstantinovskiy et al. [14]). To isolate the effect of these domain-specific trained embeddings upon the performance of our model, we run our method as described in Section 4.3 but vary the pretraining of the embeddings as follows: 1) using no embeddings at all; 2) using randomly initialized embeddings which are not pretrained; 3) using general purpose embeddings pretrained with word2vec on Google News[7]; 4) using our pretrained domain

[5]The syntactic dependency parser is available at https://spacy.io/

[7]https://code.google.com/archive/p/word2vec/

Table 2: Effectiveness of sentence check-worthiness ranking without and with weak supervision. ▲ marks statistically significant improvements with respect to the overall best baseline at the 0.05 level using a paired two tailed t-test. △ marks statistically significant improvements with respect to the best overall approach without weak supervision at the 0.05 level using a paired two tailed t-test. Significance comparisons are done on the average metric-wise performance in each of the 5 repeated runs.

	Without Weak Supervision					With Weak Supervision				
	MAP	P@5	P@10	P@20	P@R	MAP	P@5	P@10	P@20	P@R
ClaimBuster API[6]	0.230	0.219	0.176	0.159	0.138	-	-	-	-	-
TATHYA [19]	0.136	0.072	0.062	0.074	0.039	0.147	0.061	0.047	0.060	0.043
ClaimBuster [9]	0.176	0.170	0.112	0.105	0.078	0.224	0.198	0.147	0.138	0.121
Zou et al. [26]	0.187	0.143	0.105	0.099	0.086	0.212	0.171	0.111	0.121	0.097
Gencheva et al. [7]	0.238	0.276	0.170	0.153	0.123	0.236	0.222	0.143	0.138	0.113
Konstantinovskiy et al. [14]	0.267	**0.314**	0.186	0.178	0.149	0.233	0.220	0.146	0.142	0.113
Our model	**0.278**	0.291	**0.194**	**0.193**	**0.159**	**0.302**$^{▲△}$	**0.344**$^{▲}$	**0.238**$^{▲△}$	**0.218**$^{▲△}$	**0.189**$^{▲△}$

Table 3: Average overlap of syntactic dependency tags and its standard deviation between three types of sentence pairs.

Sentence pairs	Average Overlap	Standard deviation
Check-worthy	7.00	1.03
Non-check-worthy	4.74	1.08
Mixed	5.64	2.87

Table 4: Performance per type of embedding pretraining. The last row shows the performance without the syntactic dependency parsing.

Emb. pretrain	MAP	P@5	P@10	P@20	P@R
No embed.	0.184	0.153	0.116	0.103	0.087
No pretraining	0.237	0.230	0.156	0.148	0.130
Google News	0.268	0.262	0.178	0.184	0.143
Politics	**0.302**	**0.344**	**0.238**	**0.218**	**0.189**
Politics w/o syn. dep.	0.285	0.290	0.209	0.202	0.167

specific embeddings as described in Section 4.2. Table 4 shows the results when varying the embedding strategy. We see that domain specific embeddings (Politics) obtains large improvements – compared to the general purpose embedding – with improvements up to +12-34%. The last row of Table 4 shows the performance without the syntactic dependency parsing (i.e., only the word embedding), which highlights the large performance increase from the syntactic dependency parsing. As expected, no pretraining of the embeddings leads to much lower performance, though MAP is still comparable to most baselines, except for Konstantinovskiy et al. [14]. Not using embeddings at all severely drops overall effectiveness. Collectively these findings show that performance benefits more from training embeddings on smaller, yet domain-specific, data than on much larger but general domain data.

5.3 Effect of varying the amount of weakly labelled data

We analyse how our model, when used with weak supervision, scales with the amount of weakly labelled data, by reporting performance across the range of 0% to 100% weakly labelled data with

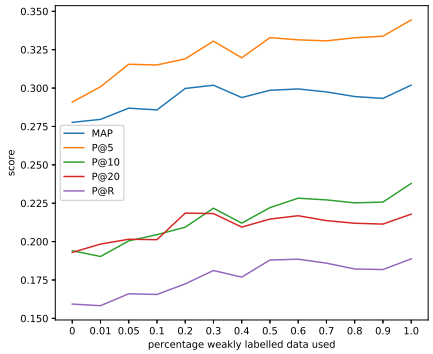


Figure 3: Impact of the amount of weakly labelled data upon our model (0 corresponds to no weakly labelled data).

10% increments. At each step we repeat the experiment 5 times with randomly sampled weakly labelled data, and report the average score. Figure 3 displays performance as a function of the percentage of weakly labelled data. As expected, the scores of all metrics generally increase as the amount of data increases. However, the largest increases happen in the first 50% of the data, and then small increases or stagnation for the remaining range up to around 90%. The performance drop at 40% may be explained by the limited number of repetitions of the sampling process, which was done due to runtime considerations. We expect that a larger number of repetitions would smooth out this slight drop.

5.4 Model interpretability

Check-worthiness detection can be part of semi-automatic or even fully manual fact checking processes, to filter claims that human fact checkers should investigate. In such cases, the output of check-worthiness detection should be easily interpretable by humans. Our model, despite being a deep learning model (generally considered to have low interpretability [24]) – provides easily interpretable output to humans through the attention mechanism that is computed on a word level (see Section 3.2). This score can be used to highlight which parts of a sentence are important for the prediction of check-worthiness. Table 5 illustrates this with a sample of true and false predictions made by our model. We see that sentences with high

predicted scores (both correctly and incorrectly predicted as check-worthy) contain a quantifiable fact consisting of a relative large number, e.g. a large amount of money (*trillion dollars, $800 billion*), a high percentage (*60 percent*), or a large collection of entities (*nearly all other presidents combined*). Sentences with low predicted check-worthiness are more varied, but generally either lack a quantifiable element or are generally vague (*buy American and hire American, fix the system, no patience for injustice*). We can also use the model to find seemingly incorrectly labeled sentences, as e.g. the non-check-worthy labelled sentences with high predictions could indeed be labelled as check-worthy instead, e.g. *"our trade deficit in goods with world last year was nearly $800 billion dollars"*.

Table 5: Check-worthy and non-check-worthy samples with high and low predictions ($\tilde{Y}$) and ground truth labels (Y). Words are colored according to attention weight: the deeper the shade of red, the larger the attention score assigned.

Y	$\tilde{Y}$	Sentence
1	0.96	america has spent approximately six trillion dollars in the middle east , all this while our infrastructure at home is crumbling .
1	0.95	today , our total business tax rate is 60 percent higher than our average foreign competitor in the developed world .
1	0.26	its the same reason why she wo nt take responsibility for her central role in unleashing isis all over the world .
1	0.22	we will follow two simple rules ; buy american and hire american .
0	0.04	millions of democrats will join our movement , because we are going to fix the system so it works fairly and justly for all americans .
0	0.05	i have no patience for injustice .
0	0.94	in the last eight years , the past administration has put on more new debt than nearly all of the other presidents combined .
0	0.95	our trade deficit in goods with the world last year was nearly $ 800 billion dollars .

6 CONCLUSION

We have presented an end-to-end trainable neural network model for ranking check-worthy sentences. The model is pretrained via weak supervision from a large collection of unlabelled data and employs a recurrent neural network with a double representation of each word using domain specific word embeddings and the syntactic dependency parsing of a sentence. We evaluate our model on check-worthy annotated political speeches from the U.S. 2016 presidential election (following the same setting as in the official CLEF 2018 competition on check-worthiness detection [17] but using even more data). Our model does not make use of specialized hand-crafted features as most related work [7, 8, 10, 19], but instead adapts the model and its representation to the domain by being trained in an end-to-end fashion. Thus, our model should by design be able to adapt to other check-worthiness tasks with results depending on the type of discourse and rhetoric. Our model effectively incorporates weak supervision: using an existing check-worthiness ranking system to weakly label political speeches significantly improved performance. Overall, our model outperforms all state of the art baselines in mean average precision and precision at different cut offs, with statistically significant +9-28% gains from the best performing baseline. Further analysis revealed the significance of domain specific word embeddings, compared to traditional general purpose embeddings, and how check-worthy sentences share a syntactic similar structure than non-check-worthy sentences.

Future work consists of investigating further multiple weak signals and incorporating text discourse context into the model.

ACKNOWLEDGMENTS

Partly funded by Innovationsfonden DK, DABAI (5153-00004A), and AMAOS (7076-00121B).

REFERENCES

[1] Mouhamadou Lamine Ba, Laure Berti-Equille, Kushal Shah, and Hossam M Hammady. 2016. Vera: A platform for veracity estimation over web data. In *International Conference Companion on World Wide Web*. 159–162.
[2] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 33–40.
[3] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Syntax, Semantics and Structure in Statistical Translation* (2014), 103–111.
[4] Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Annual Meeting of the Association for Computational Linguistics*. 387–396.
[5] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015).
[6] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Conference on Empirical Methods in Natural Language Processing*. 670–680.
[7] Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In *International Conference Recent Advances in Natural Language Processing*. 267–276.
[8] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1803–1812.
[9] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, et al. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.
[10] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. In *Conference of the North American Chapter of the Association for Computational Linguistics*. 26–30.
[11] Damian Jimenez and Chengkai Li. 2018. An Empirical Study on Identifying Sentences with Salient Factual Statements. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[12] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully Automated Fact Checking Using External Sources. In *International Conference Recent Advances in Natural Language Processing*. 344–353.
[13] Diane Kelly, Xin Fu, and Chirag Shah. 2010. Effects of position and number of relevant documents retrieved on users' evaluations of system performance. *Transactions on Information Systems (TOIS)* 28, 2 (2010), 9.
[14] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. http://arxiv.org/abs/1809.08193
[15] Dieu-Thu Le, Ngoc Thang Vu, and André Blessing. 2016. Towards a text analysis system for political debates. In *LaTeCH@ACL*.
[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[17] Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, et al. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In *International Conference of the CLEF Association*.

[18] Yifan Nie, Alessandro Sordoni, and Jian-Yun Nie. 2018. Multi-level abstraction convolutional model with weak supervision for information retrieval. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*. 985–988.

[19] Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. TATHYA: A multi-classifier system for detecting check-worthy statements in political debates. In *ACM on Conference on Information and Knowledge Management*. 2259–2262.

[20] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, et al. 2014. A Gold Standard Dependency Corpus for English. In *International Conference on Language Resources and Evaluation*. 2897–2904.

[21] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3346–3359.

[22] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[23] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 105–114.

[24] Quanshi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of IT & EE* 19, 1 (2018), 27–39.

[25] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 32.

[26] Chaoyuan Zuo, Ayla Ida Karakas, and Ritwik Banerjee. 2018. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In *CLEF 2018 Working Notes* (10 ed.). CEUR-WS.org.

Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-checking

Naeemul Hassan
Computer and Information Science
University of Mississippi
nhassan@olemiss.edu

Mohammad Yousuf
Journalism and Mass Communication
University of Oklahoma
yousuf@ou.edu

Md Mahfuzul Haque
Journalism and New Media
University of Mississippi
mhaque@go.olemiss.edu

Javier A. Suarez Rivas
Computer and Information Science
University of Mississippi
jsuarez@go.olemiss.edu

Md Khadimul Islam
Political Science
University of Mississippi
mislam1@go.olemiss.edu

ABSTRACT

This study explores an online fact-checking community called *politicalfactchecking* on reddit.com that relies on crowdsourcing to find and verify check-worthy facts relating to U.S. politics. The community embodies a network journalism model in which the process of finding and verifying check-worthy facts through crowdsourcing is coordinated by a team of moderators. Applying the concepts of connective journalism, this study analyzed the posts ($N = 543$) and comments ($N = 10,221$) on the community's Reddit page to understand differences in roles of the community members and the moderators. A mixed-method approach was used to analyze the data. The authors also developed an automated argument classification model to analyze the contents and identify ways to automate parts of the process. The findings suggest that a model consisting of crowds, professionals, and computer-assisted analysis could increase efficiency and decrease costs in news organizations that involve fact-checking.

CCS CONCEPTS

• **Human-centered computing → Collaborative content creation**; *Collaborative interaction.*

ACM Reference Format:
Naeemul Hassan, Mohammad Yousuf, Md Mahfuzul Haque, Javier A. Suarez Rivas, and Md Khadimul Islam. 2019. Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-checking. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316734

1 MOTIVATION

Many professional fact-checkers are suspicious of the idea of crowdsourced fact-checking in which users verify factual claims [7]. Professionals often claim that users lack required skills, and are biased. We argue that contribution of crowds to fact-checking is essential in the networked media ecosystem where information is abundant and rumors spread like wildfire with resources for investigative journalism steadily plummeting [6, 14]. A sustainable model for fact-checking platforms would consist of crowds, professional fact-checkers, and automated assisting tools. Crowds can perform many mundane but important tasks under the guidance of professionals while programmers build tools to find credible sources and make sense of a large amount of user inputs. Crowds can also help better identify facts that people care about and identify sources that may lack credibility but still are popular. Professionals can play the 'moderator' and 'seminar leader' role in this process [16]. The purpose of this study is to explore such a crowdsourced fact-checking model where users check facts under the guidance of moderators. This study adds 'automation' to that model and identifies roles of each component of this model. It examines posts and comments in a crowd curated political fact-checking platform of *reddit.com/politicalfactchecking* to find answers to the following questions-

- What claims do people want to see fact-checked?
- What role does crowd play in fact-checking a claim?
- What roles do moderators play in crowdsourced fact-checking?
- How can computation play a role in scaling-up crowdsourced fact-checking?

To understand roles of the crowd in finding factual claims, we analyzed the posts in *reddit/politicalfactcheck* and identified popular topics that drew more user participation (e.g., number of comments), original sources of claims (e.g., mainstream media, alternative media, blogs), and type of claims (e.g., stat, figure). To understand the roles of crowd in fact-checking, we analyzed the comments and identified types of arguments users provide, originality of sources that users present to support or reject claims, and relation between crowd actions and professional fact-checking activities. To identify roles of moderators, we conducted a qualitative analysis on all posts and comments of moderators. We've also developed an automated argument classification model and identified some steps in fact-checking, which could be automated.

The findings suggest *Reddit's* model can lay a foundation for building a sustainable model for fact-checking organizations that lack resources. Roles of the crowd and automated assisting tools in fact-checking can increase efficiency and decrease expense and thus will enhance sustainability. Models solely reliant on professionals

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316734

are not sustainable given the amount of resource they require to survive.

2 RELATED WORK

Professionals and scholars described fact-checking as a complicated process consisting of various steps that can be grouped into three major categories–(i) selection of facts to check, (ii) collection of evidence, and (iii) decision [3, 5, 7]. Each of these steps comprises of multiple sub-steps. For instance, fact selection includes choosing claims "from countless public utterances", separating facts from opinion, separating newsworthy/check-worthy facts from all facts, and filtering verifiable facts [7]. In the literature on news media, fact-checking is narrowly defined and refers to as verification[5]. It is believed to be a job of trained professionals. Graves noted that professional fact-checkers practice a type of "intertextual and annotative journalism" that lies "within the framework of a larger political critique" [7]. Fact-checking sites run by professionals "operate very self-consciously as hybrids of old and new media practice, organizing veteran print and broadcast reporters around a modernizing genre meant to update political journalism" [7].

With the advent of digital technologies, research on fact-checking started to broaden its horizon. Mark Little [15] asked journalists to "get comfortable with risk, transparency and collaboration" and find "the wisdom in the crowd". Cohen et al. [2] envisioned a system called a *cloud for the crowd*, which combines computational resources as well as human expertise to support more efficient and effective investigative journalism. In a recent white paper [1], Mevan et al. presented the state of automated fact-checking and described how fact-checking can be scaled up dramatically using existing technologies. Hassan et al. suggested a *Holy Grail* towards computational fact-checking by depicting the roles of professional fact checkers and automation [8]. They also studied how automated claim detection techniques perform against professional fact-checkers [11]. A number of tools have recently been proposed to fully automate fact-checking using natural language processing, machine learning, knowledge graph query, and question-answering techniques [9, 10, 12, 20]. However, there is a lack of research on studying the active role of crowd in fact-checking. According to the best of our knowledge, this is the first work which systematically studies fact-checking in a crowdsourced environment.

3 METHODOLOGY

3.1 Dataset

In this work, we study the crowd curated online political fact-checking platform *reddit/politicalfactchecking* [2]. By and large, the platform functions in the following way– **i)** a user creates a post to submit a claim, **ii)** the community presents evidence and arguments related to the claim using the platform's comment and reply features, **iii)** a group of moderators decides the appropriate flair for the claim based on the accumulated evidence and provides a justification. Note, both the post creator and the moderator group are members of the community and are eligible to participate in the second phase. A claim's flair can be modified later in case there is new evidence. Below, we provide a short description of each of the flairs. Detailed explanations are available in the platform's website. We say a claim is *fact-checked* if the corresponding post is assigned any of the flairs except *Please Verify* and *Mod Post/Meta*.

Confirmed: This post has enough evidence to support the claim.
Mostly True: The claim is accurate but needs clarification/context.
Half True: The claim is not entirely accurate, leaves out important information, or is out of context.
Mostly False: The claim contains some elements of truth but ignores critical facts that would change the reader's impression.
False: The claim is blatantly false.
Partisan Bias: This claim contains obvious political bias intended to edify one party or make another party look bad.
Unverifiable: This claim contains more opinion than fact. It can easily become a debate. This tag also covers doublespeak, instances of incorrect terminology, or other miscellaneous claims that cannot be proved true or false.
Please Verify: This post needs to be researched more in order to determine the validity of the claim.
Mod Post/Meta: This is a post by a moderator or a discussion about improving *reddit/politicalfactchecking*.

Table 1: Distribution of the flairs

Flair	# Posts	Avg. # Comments+Replies per Post
Confirmed	32	20
Mostly True	21	23
Half-True	18	27
Mostly False	13	18
False	70	23
Unverifiable	39	27
Partisan Bias	16	18
Please Verify	111	18
Mod Post/Meta	44	43
No Flair	179	13

Using PRAW (Python Reddit API Wrapper) [3], a Python package which allows access to *reddit's* official API, we scraped all (everything before July, 2017) the posts, comments and their metadata (timestamp, author). In total, there are 543 posts, 2,835 comments, and 7,386 replies to the comments. Table 1 shows number of posts and the average number of comments and replies per post for each flair category. There are 209 posts which have been fact-checked, 111 posts in *Please Verify* category, 179 posts which didn't receive any flair, and 44 posts in *Mod Post/Meta* category.

Figure 1 shows number of posts created in each quarter since the platform's creation in September 2012 till July, 2017. It appears that the platform was more active at the initial stage. There were 10,083 members, 2,019 unique commentators, and 5 moderators in the community. About 80% of the comments and replies were created by 28% of the commentators. The average lengths of comments and replies are 477 and 331, respectively.

3.2 Coding and Analysis

We manually coded the posts and comments to answer research questions. The posts were coded for five variables: *post topic, post*

[1] https://fullfact.org/blog/2016/aug/automated-factchecking/
[2] https://www.reddit.com/r/politicalfactchecking
[3] https://praw.readthedocs.io/en/latest/

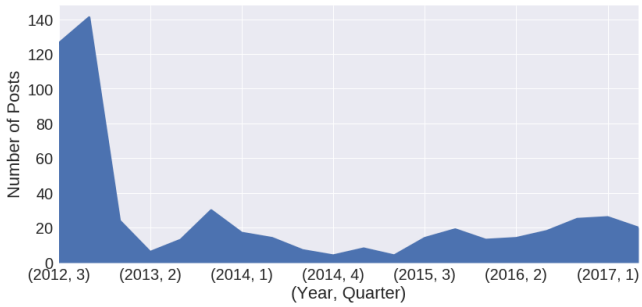


Figure 1: Number of posts per quarter

type, source of claim, nature of the claim, and *fact-checked entity*. The comments were coded for three variables: *commenter action, argument type*, and *source of evidence*. We used a combination of analytical approaches to identify these variables and categories, as we did not find any study that analyzed *Reddit* posts for the purpose of understanding roles of crowds in fact-checking. We used a combination of deductive and inductive approaches [18]. We initially applied a deductive approach to identify the variables and the categories through a review of professional and scholarly literature on media content and typologies [13, 21]. Then, we used an inductive approach to refine those categories to fit the purpose of this study. The authors have had several training sessions to discuss and finalize the coding categories. One coder coded most of the posts while the comments were coded by four coders, three of whom were journalists with over 30 years of experience combined. Inter-coder reliability (Krippendorff's α)[4] ranged between 0.92 and 0.96 for comments, with an average of 0.94.

3.2.1 Posts. The first variable, *post topic*, included 12 categories: *Economy*, *Education*, *Election*, *Environment*, *Foreign Affairs*, *Health*, *Immigration*, *Equality*, *Media Bias*, *Religion*, *Security*, and *Other*. The second variable, *post type*, included three categories: *Fact-check Request*, *Seeking Information*, and *Judgement/Opinion*. We identify six categories for *source of claim*: *News/Information Media*, *Organization*, *Alternative Media*–websites where majority of contents is created by contributors, *User Generated Content (UGC) Platforms* (e.g., social media), *Personal Contact*, and *Others*. The fourth variable, *nature of the claim*, was coded for four categories: *Non-statistical* (e.g., quotes), *Statistical*, *Media* (e.g., photo or video), and *Other*. The fifth variable, *fact-checked entity*, included *Person*, *Organization*, *Event*, *Policy/Issue*, and *Other*. Note that a post may be coded with multiple categories for a variable.

3.2.2 Comments. The same categories used for the variable *post type* and *source of claim* in posts were used for the variable *argument type* and *source of evidence*, respectively. The *commenter action* variable was coded for six categories: *Providing Argument*, *Seek Clarification*, *Check Verifiability*, *Contact Claim Source*, *Assign Flair*, and *Post Irrelevant*.

3.2.3 Qualitative Analysis. To understand the roles of the moderators, we qualitatively analyzed the posts and comments published by the moderators. We applied an approach developed by Altheide et al.[1], which is widely used in various fields.

4 RESULTS AND DISCUSSION

Following the above explained methods, we analyze the *political-factchecking* dataset, present the results and discuss the findings in this section.

4.1 Role of Crowd in Finding Claims

There are 499 posts in the dataset published by the crowd. Table 2 shows topic distribution of the claims. 70% of all the claims belong to the top 4 topics– *Economy*, *Election*, *Security*, *Health*. For each topic, we identified the posts which have been fact-checked (*Confirmed*, *Mostly True*, *Half True*, *Mostly False*, *False*, *Unverifiable*, *Partisan Bias*) and which have not been fact-checked (*Please Verify*, *No Flair*). Table 2 shows number of fact-checked posts, number of not fact-checked posts, and the percentage of fact-checked posts for each topic. Among the topics having at least 10 posts, *Health*, *Foreign Affairs*, and *Economy* related posts have significantly higher probability of getting fact-checked than the *Immigration* and *Equality* related posts.

Among all the posts, 352 of them are request for fact-checking claims, 108 are about seeking opinion/information (e.g., *where can I find ...*), and 49 posts are giving judgment regarding an issue without requesting any fact-check. About half of the fact-checking requests are statistical in nature (e.g., check a number, check degree of a phenomenon) and other half are non-statistical (e.g., check if someone said something). There are 13 requests for checking authenticity of photos and videos.

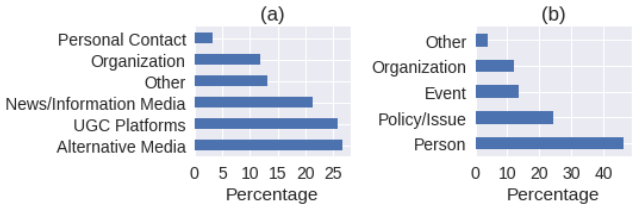


Figure 2: Distribution of source categories (a) and checked entities (b)

We investigated the sources of the claims which crowd wants to see fact-checked. Figure 2a shows the distribution of source types. We observe that *Alternative Media* and *UGC Platform* are the major categories of sources the crowd wants to fact-check. These two source categories together cover 53% of all the claims. The most common UGC platform is *imgur.com* [4]. We also study what entities the crowd wants to fact-check. Most (45%) of the posted claims are about *Person* (e.g., Barack Obama, Donald Trump) and *Policy/Issue* (e.g., Medicare, Tax rate).

4.2 Role of Crowd in Checking Claims

Fact-checking is a complex task which consists of multiple sub-processes. It's interesting to see how the crowd self-organize itself to accomplish this task by contributing to the subprocesses. We analyzed the comments and replies to study the different types of actions performed by the crowd. Table 3 shows the distribution of actions in comments. Note that a comment can be long and may contain multiple actions. In such cases, we labeled it with multiple

[4] http://imgur.com/

Table 2: Distribution of topics

Topic	# Posts	# Fact-checked	# Not Fact-checked	% Fact-checked
Economy	130	62	68	47.69
Election	83	34	49	40.96
Security	68	24	44	35.29
Health	63	30	33	47.62
Other	39	12	27	30.77
Foreign Affairs	34	16	18	47.06
Media Bias	27	10	17	37.04
Equality	20	6	14	30.00
Immigration	14	3	11	21.43
Religion	6	4	2	66.67
Education	5	4	1	80.00
Environment	4	2	2	50.00

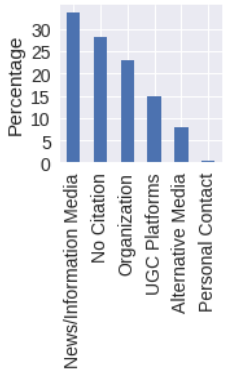


Figure 3: Source category distribution

actions. That is why the percentages do not sum up to hundred. About (70%) of the comments are used to provide argument. The crowd also identifies if a posted claim is verifiable or not and seek for information if further clarification is required. Some users also post irrelevant (ads, jokes) comments. We observe that contacting source of a claim, a primary task in professional fact-checking, is not practiced by the crowd. Only one comment was found where the user mentioned contacting the source.

We further studied the nature of the provided arguments. There are 1,080 arguments which contain factual evidence. 701 of these contain additional information leading to a fact-supported inference, either supporting the claim or refuting it. We observe that some commenters give judgment or opinion without factual justification. There are 860 such cases. We also analyzed the source of the

Table 3: Distribution of actions performed by the crowd

Action	# Comments	Percentage
Provide Argument	1942	69.56
Post Irrelevant	402	14.40
Seek Clarification	355	12.71
Assign Flair	141	5.05
Check Verifiability	48	1.72
Contact Claim Source	1	0.04

Table 4: Top-12 cited domains

Domain	# Citations
wikipedia.org	298
politifact.com	150
reddit.com	111
washingtonpost.com	91
nytimes.com	78
youtube.com	68
whitehouse.gov	51
factcheck.org	40
huffingtonpost.com	34
imgur.com	32
forbes.com	29
cnn.com	27

Table 5: Average number of factual evidence and opinionated judgments per flair

Flair	Avg. # Factual Evidence	Avg. # Opinion/Judgment
Mostly True	3.22	1.83
Half True	3.17	1.67
Confirmed	3.10	1.56
False	2.89	2.59
Mostly False	2.75	1.86
Partisan Bias	2.73	1.90
Please Verify	2.59	2.40
Unverifiable	2.38	3.33
None	1.74	3.10

factual evidence provided by the crowd. Figure 3 shows distributions of the categories. The top category is *News/Information Media*. There are 305 cases (28%) where the commenter presented factual evidence but didn't provide citation. Table 4 shows the top-12 cited domains. In many cases, the commenters cite evidence from popular fact-checking platforms such as *politifact.com* [5] and *factcheck.org* [6]. Table 5 shows the average number of comments containing factual evidence per post for each flair category. It is observed that, in general, *Unverifiable*, *Please Verify* flair categories receive less factual evidence and more opinionated comments compared to the conclusive (e.g., *Confirmed*, *Mostly True*) categories.

We measured the number of days required for the crowd to fact-check a claim. The *Reddit* API doesn't provide the exact flair assignment timestamp. A simple heuristic gives us an approximation of the flair assignment timestamp. We observe that before assigning a flair to a post, the moderators publish a comment mentioning and justifying the flair to be assigned. We use the timestamp of the latest comment made by a moderator containing the assigned flair's words in the text as an approximation of the flair assignment timestamp. Subtracting the post creation timestamp from this gives us an approximation of the required fact-checking time. The mean, median and standard deviation of the required fact-checking time is 7, 1, and 20 days, respectively.

[5] http://www.politifact.com/
[6] http://www.factcheck.org/

4.3 Role of Moderators

A qualitative analysis [1] of the moderator posts revealed several roles that moderators play to keep discussions on topic and fact-based. At the submission level, a moderator's roles can be compared to those of a gatekeeper who defines rules of a group and decides which posts qualify for fact-checking. At the comment level, moderators acted as "seminar leaders" in which they took part in verification, analysis, and evaluation of evidence posted by users.

The most prominent role of moderators at the submission level, defined by the frequency of posts with similar themes, emerged as reminding users of the rules and regulations, and issuing warnings. For instance, one post starts as:

"*I've seen posts in this thread increasingly devolving into political bickering, off-topic arguing, and.....Get your act together, keep things on-topic, and keep things fact-based. Otherwise, your posts or comments may be removed".*

The second most prominent role of moderator is to encourage users by thanking them for participation and showcasing impact of their works. For instance, part of a post reads:

"*so many of you make this a fantastic little sub with a lot to offer. Please keep it up.*"

Another post adds, "*...Presidential debate drew over 16,000 page views here and more than 4,330 unique viewers.*"

The third prominent role is to announce new events and topics to be covered. Other roles include seeking suggestions on various topics (e.g., how to cover a live debate). Some of the moderator posts were short while others provided more details. The number of words in a post ranges between one and 673.

The moderators posted 381 (13.44%) out of 2, 835 comments. In addition to assigning flairs based on available evidence, moderators took on several roles at the comment level. We find 165 instances where the moderators themselves presented source with factual evidence. Such sources include links to books, columns, datasets, news articles, press releases, research papers, and transcripts. The following comment represents a good example of how moderators help users:

"*OK one more source. This from another biography on Michelle, this link is super long so i will post that here then the quote in a reply*".

One noticeable role of moderators appeared to be changing flair as more evidence is presented. For instance, one moderator had labeled a fact as mostly true:

"Marking this as Mostly True. Yes, 4.5M jobs were created, since the lowest point of the recession. a net gain of 300K since the start of Obama's administration."

As more evidence came in, another moderator changed the flair to half true:

"Marking this one as 'Half True' based on everyone's research here. The 4.5 million jobs created seems factually verified. Whether it was better than the Bush recovery from the first recession seems unclear. But Cutter seems to have suggested she was in error on the Reagan recovery.".

Another role was to ask for clarification about vague posts. For example, a typical moderator comment representing this role reads:

"*So can you help us with what part of this you'd like us to fact check? The US military quote from a NY Times article as accurate? Or other?*" We find 69 similar comments.

4.4 Role of Computation

After observing the way *politicalfactchecking* operates, we identified several places where computation techniques can work with the crowd hand in hand and produce a faster, scalable and sustainable fact-checking model.

Argument Classification: One critical role of the moderators is to go through the arguments provided by the crowd and produce a conclusive decision based on the evidence. We argue, machine learning techniques can facilitate this task by automatically separating factual evidence from opinionated or evidence lacking judgment. Even though the dataset at hand is small in size, nonetheless, we built a binary classifier using Gradient Boosting algorithm and trained it over the manually coded comments (coding details in section 3.2.2 and 4.2). Word tokens were used as features after removing the stop words. The model was evaluated using 5-fold cross-validation. Its average precision of identifying comments containing *factual evidence* and *evidence lacking judgment* is 85% and 82%, respectively (i.e., the model is accurate 85% of the times when it says that a comment contains factual evidence). In a crowdsourced environment, where the number of moderators is significantly smaller than the crowd size, such an automated argument classifier can greatly reduce the workload of the moderators by straining the evidence lacking judgments from large number of comments.

Stance Detection: Another important task in fact-checking is to juxtapose supporting and opposing evidence. The moderators scrutinize the comments and identify the supporting/opposing evidences. We argue, like the argument classification task, identifying stance of an evidence can also be automated using computation techniques; particularly, natural language processing, computational reasoning, and machine learning. Given a claim and an evidence, the goal should be automatically detecting the stance of the evidence towards the claim. Our preliminary investigation over the *reddit/politicalfactchecking* dataset suggests that stance detection is a harder problem than argument classification. [17, 19] have studied the stance detection problem where the target is an issue rather than a claim. Recently, *Fake News Challenge*[7] have released a large dataset of news headline-body pairs with manually labeled stance information. We plan to continue investigating the stance detection problem.

In addition to the above mentioned tasks, there are other rooms for automation as well. For instance, given a claim, identify existing fact-checks or related evidence from the web, predicting the veracity of a claim based on provided evidence, soliciting arguments from users based on expertise, and so on.

5 CONCLUSION AND FUTURE WORK

Despite of many limitations of crowdsourced fact-checking on *reddit/politicalfactchecking*, this model outlived many fact-checking

[7] http://www.fakenewschallenge.org/

sites [8]. It uses less resources compared to most professional fact-checking sites and utilizes the wisdom of the crowd to check hundreds of facts. Yet, it received negligible attention from scholars and professionals. Our findings show strong potential of this crowdsourced fact-checking model. The results contradict traditional beliefs that crowd is unable to fact-check and they provide opinions, not facts [9]. We found that more than 50% of the user comments providing an argument in support or against a claim had contained factual evidence. This contradiction may result from the contribution of moderators who often intervene to keep discussion on topic and fact-based. This study has revealed that crowds do more than fact-checking. They, among other roles, help find facts, play a role in deciding which facts are worth checking as well as which facts are verifiable. What professional fact-checking sites can learn from this study is that a small number of trained fact-checkers can help check a large number of facts if they can properly lead crowds in the right direction.

This study also adds to previous research on automated fact-checking tools which could strengthen the crowdsourced model and improved its efficiency in multitudes. In conclusion, a future fact-checking model must incorporate all three components–automated assisting tools, crowds and professionals–to survive and thrive in the current media ecosystem. Each of these components complements one another. In future, we plan to investigate the comment replies and understand how discussion forms in a crowdsourced fact-checking model. We also intend to build robust stance detection techniques to facilitate the fact-checking process.

REFERENCES

[1] David L. Altheide and Christopher J. Schneider. 2013. *Qualitative media analysis.* Sage.

[2] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: A Call to Arms to Database Researchers.. In *CIDR*, Vol. 2011. 148–151.

[3] David A. Craig and Mohammad Yousuf. 2013. Excellence in journalistic use of social media through the eyes of social media editors. (2013).

[4] Knut De Swert. 2012. Calculating inter-coder reliability in media content analysis using Krippendorff's alpha. *Center for Politics and Communication* (2012).

[5] Michael Dobbs. 2012. *The rise of political fact-checking.* New America Foundation.

[6] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. Why rumors spread so quickly in social networks. *Commun. ACM* (2012).

[7] Lucas Graves. 2013. *Deciding what's true: Fact-checking journalism and the new ecology of news.* Columbia University.

[8] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. *Proceedings of the 2015 Computation+Journalism Symposium* (2015).

[9] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2017).

[10] Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.* ACM, 1835–1838.

[11] Naeemul Hassan, Mark Tremayne, Fatma Arslan, and Chengkai Li. 2015. Comparing automated factual claim detection against judgments of journalism organizations. *Proceedings of the 2016 Computation+Journalism Symposium* (2015).

[12] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. *Proceedings of the VLDB Endowment* 10, 7 (2017).

[13] S.I. Hayakawa and Alan R. Hayakwa. 1990. *Language in thought and action.* Houghton Mifflin Harcourt.

[14] Brant Houston. 2010. The future of investigative journalism. *Daedalus* (2010).

[15] Mark Little. 2012. Finding the wisdom in the crowd. *Nieman Reports* (2012).

[16] Wilson Lowrey and Peter J Gade. 2011. *Changing the news: The forces shaping journalism in uncertain times.* Routledge.

[17] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. [n. d.]. A Dataset for Detecting Stance in Tweets.

[18] Kimberly A. Neuendorf. 2002. *The content analysis guidebook.* Sage.

[19] Parinaz Sobhani. 2017. *Stance Detection and Analysis in Social Media.* Ph.D. Dissertation. Université d'Ottawa/University of Ottawa.

[20] Brett Walenz, Y Wu, S Song, Emre Sonmez, Eric Wu, Kevin Wu, Pankaj K Agarwal, Jun Yang, Naeemul Hassan, Afroza Sultana, et al. 2014. Finding, monitoring, and checking claims computationally based on structured data. In *Computation+Journalism Symposium.*

[21] Mohammad Yousuf. 2016. *Niches in the networked media ecosystem: Functionalities, gratifications, and a typology of online news and informational media.* Ph.D. dissertation. The University of Oklahoma.

[8] https://www.poynter.org/news/why-do-fact-checking-sites-close-and-how-can-new-ones-avoid-fate

[9] https://www.journalism.co.uk/news/could-crowdsourcing-help-you-fact-check-your-data-/s2/a557045

Red Bots Do It Better: Comparative Analysis of Social Bot Partisan Behavior

Luca Luceri*
University of Applied Sciences and Arts of Southern Switzerland, and University of Bern
Manno, Switzerland
luca.luceri@supsi.ch

Ashok Deb
USC Information Sciences Institute
Marina del Rey, CA
ashok@isi.edu

Adam Badawy
USC Information Sciences Institute
Marina del Rey, CA
badawy@isi.edu

Emilio Ferrara
USC Information Sciences Institute
Marina del Rey, CA
emiliofe@usc.edu

ABSTRACT

Recent research brought awareness of the issue of bots on social media and the significant risks of mass manipulation of public opinion in the context of political discussion. In this work, we leverage Twitter to study the discourse during the 2018 US midterm elections and analyze social bot activity and interactions with humans. We collected 2.6 million tweets for 42 days around the election day from nearly 1 million users. We use the collected tweets to answer three research questions: (*i*) *Do social bots lean and behave according to a political ideology?* (*ii*) *Can we observe different strategies among liberal and conservative bots?* (*iii*) *How effective are bot strategies in engaging humans?*

We show that social bots can be accurately classified according to their political leaning and behave accordingly. Conservative bots share most of the topics of discussion with their human counterparts, while liberal bots show less overlap and a more inflammatory attitude. We studied bot interactions with humans and observed different strategies. Finally, we measured bots embeddedness in the social network and the extent of human engagement with each group of bots. Results show that conservative bots are more deeply embedded in the social network and more effective than liberal bots at exerting influence on humans.

CCS CONCEPTS

• **Networks → Social media networks**; • **Human-centered computing → Social network analysis**.

KEYWORDS

social media; political elections; social bots; political manipulation

*Also with USC Information Sciences Institute, Marina del Rey, CA.

L. Luceri & A. Deb contributed equally to this work.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316735

ACM Reference Format:
Luca Luceri, Ashok Deb, Adam Badawy, and Emilio Ferrara. 2019. Red Bots Do It Better: Comparative Analysis of Social Bot Partisan Behavior. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316735

1 INTRODUCTION

During the last decade, social media have become the conventional communication channel to socialize, share opinions, and access the news. Accuracy, truthfulness, and authenticity of the shared content are necessary ingredients to maintain a healthy online discussion. However, in recent times, social media have been dealing with a considerable growth of false content and fake accounts. The resulting wave of misinformation (and disinformation) highlights the pitfalls of social media and their potential harms to several constituents of our society, ranging from politics to public health.

In fact, social media networks have been used for malicious purposes to a great extent [11]. Various studies raised awareness about the risk of mass manipulation of public opinion, especially in the context of political discussion. Disinformation campaigns [2, 5, 12, 14–17, 22, 24, 26, 30] and social bots [3, 4, 21, 23, 25, 29, 31, 32] have been indicated as factors contributing to social media manipulation.

The 2016 US Presidential election represents a prime example of the significant perils of mass manipulation of political discourse. Badawy et al. [1] studied the Russian interference in the election and the activity of Russian trolls on Twitter. Im et al. [18] suggested that troll accounts are still active to these days. The presence of social bots does not show any sign of decline [10, 32] despite the attempts from social network providers to suspend suspected, malicious accounts. Various research efforts have been focusing on the analysis, detection, and countermeasures development against social bots. Ferrara et al. [13] highlighted the consequences associated with bot activity in social media. The online conversation related to the 2016 US presidential election was further examined [3] to quantify the extent of social bots activity. More recently, Stella et al. [27] discussed bots' strategy of targeting influential humans to manipulate online conversation during the Catalan referendum for independence, whereas Shao et al. [25] analyzed the role of social bots in spreading articles from low credibility sources. Deb et al.

[10] focused on the 2018 US Midterms elections with the objective to find instances of voter suppression.

In this work, we investigate social bots behavior by analyzing their activity, strategy, and interactions with humans. We aim to answer the following research questions (RQs) regarding social bots behavior during the 2018 US Midterms election.

RQ1: *Do social bots lean and behave according to a political ideology?* We investigate whether social bots can be classified based on their political inclination into liberal or conservative leaning. Further, we explore to what extent they act similarly to the corresponding human counterparts.

RQ2: *Can we observe different strategies among liberal and conservative bots?* We examine the differences between social bot strategies to mimic humans and infiltrate political discussion. For this purpose, we measure bot activity in terms of volume and frequency of posts, interactions with humans, and embeddedness in the social network.

RQ3: *How effective are bot strategies in engaging humans?* We introduce four metrics to estimate the effectiveness of bot strategies in involving humans in their conversation and to evaluate the degree of human interplay with social bots.

We leverage Twitter to capture the political discourse during the 2018 US midterm elections. We collected 2.6 million tweets for 42 days around election day from nearly 1 million users. We then explore collected data and attain the following findings:

- We show that social bots are embedded in each political side and behave accordingly. Conservative bots abide by the topic discussed by the human counterpart more than liberal bots, which in turn exhibit a more provocative attitude.
- We examined bots' interactions with humans and observed different strategies. Conservative bots stand in a more central social network position, and divide their interactions between humans and other conservative bots, whereas liberal bots focused mainly on the interplay with the human counterparts.
- We measured the extent of human engagement with bots and recognized the strategy of conservative bots as the most effective in terms of influence exerted on human users.

2 DATA

In this study, we use Twitter to investigate the partisan behavior of malicious accounts during the 2018 US midterm elections. For this purpose, we carried out a data collection from the month prior (October 6, 2018) to two weeks after (November 19, 2018) the day of the election. We kept the collection running after the election day as several races remained unresolved. We employed the Python module *Twyton* to collect tweets through the Twitter Streaming API using the following keywords as a filter: *2018midtermelections*, *2018midterms*, *elections*, *midterm*, and *midtermelections*. As a result, we gathered 2.7 million tweets, whose IDs are publicly available for download.[1] From this set, we first removed any duplicate tweet, which may have been captured by accidental redundant queries to the Twitter API. Then,we filtered out all the tweets not written in English language and those that were out of the context of this study. Overall, we retain nearly 2.6 millions tweets, whose aggregate statistics are reported in Table 1.

Table 1: Dataset statistics

Statistic	Count
# of Tweets	452,288
# of Retweets	1,869,313
# of Replies	267,973
# of Users	997,406

3 METHODOLOGY

3.1 Bot Detection

Nowadays, bot detection is a fundamental asset for understanding social media manipulation and, more specifically, to reveal malicious accounts. In the last few years, the problem of detecting automated accounts gathered both attention and concern [13], also bringing a wide variety of approaches to the table [7, 8, 20, 28]. While increasingly sophisticated techniques keep emerging [20], in this study, we employ the widely used *Botometer*.[2]

Botometer is a machine learning-based tool developed by Indiana University [9, 29] to detect social bots in Twitter. It is based on an ensemble classifier [6] that aims to provide an indicator, namely *bot score*, used to classify an account either as a bot or as a human. To feed the classifier, the Botometer API extracts about 1,200 features related to the Twitter account under analysis. These features fall in six broad categories and characterize the account's profile, friends, social network, temporal activity patterns, language, and sentiment. Botometer outputs a bot score: the lower the score, the higher the probability that the user is human. In this study we use version *v3* of Botometer, which brings some innovations, as detailed in [32]. Most importantly, the bot scores are now rescaled (and not centered around 0.5 anymore) through a non-linear re-calibration of the model.

In Figure 1, we depict the bot score distribution of the 997,406 distinct users in our datasets. The distribution exhibits a right skew: most of the probability mass is in the range [0, 0.2] and some peaks can be noticed around 0.3. Prior studies used the 0.5 threshold to separate humans from bots. However, according to the re-calibration introduced in Botometer *v3* [32], along with the emergence of increasingly more sophisticated bots, we here lower the bot score threshold to 0.3 (i.e., a user is labeled as a bot if the score is above 0.3). This threshold corresponds to the same level of sensitivity setting of 0.5 in prior versions of Botometer (cf. Fig 5 from [32]).

According to this choice, we classified 21.1% of the accounts as bots, which in turn generated 30.6% of the tweets in our data set. Overall, Botometer did not return a score for 35,029 users that corresponds to 3.5% of the accounts. We used the Twitter API to further inspect them. Interestingly, 99.4% of these accounts were suspended by Twitter, whereas the remaining percentage of users protected their tweets turning on the privacy settings of their accounts.

3.2 Political Ideology Inference

In parallel to the bot detection analysis, we examine the political leaning of both bots and humans in our dataset. To classify users based on their political ideology, we rely on the political leaning of

[1] https://github.com/A-Deb/midterms

[2] https://botometer.iuni.iu.edu/

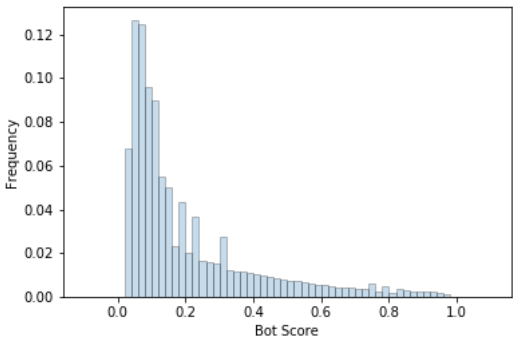


Figure 1: Bot score distribution

the media outlets they share. We make use of a list of partisan media outlets released by third-party organizations, such as AllSides[3] and Media Bias/Fact Check.[4] We combine liberal and liberal-center media outlets into one list (composed of 641 outlets) and conservative and conservative-center into another (composed of 398 outlets). To cross reference these media URLs with the URLs in the Twitter dataset, we need to get the expanded URLs for most of the links in the dataset, as most of them are shortened. However, this process is quite time-consuming, thus, we decided to rank the top 5,000 URLs by popularity and retrieve the long version only for those. These top 5,000 URLs accounts for more than 254K, or more than 1/3 of all the URLs in the dataset. After cross-referencing the 5,000 extended URLs with the media URLs, we observe that 32,115 tweets in the dataset contain a URL that points to one of the liberal media outlets and 25,273 tweets with a URL pointing to one of the conservative media outlets.

To label Twitter accounts as liberal or conservative, we use a polarity rule based on the number of tweets they produce with links to liberal or conservative sources. Thereby, if an account has more tweets with URLs pointing to liberal sources, it is labeled as liberal and vice versa. Although the overwhelming majority of accounts include URLs that are either liberal or conservative, we remove any account that has equal number of tweets from each side. Our final set of labeled accounts includes 38,920 users.

Finally, we use label propagation to classify the remaining accounts in a similar way to previous work (*cf.* [1]). For this purpose, we construct a social network based on the retweets exchanged between users. The nodes of the retweet network are the users, which are connected by a direct link if one user retweeted a post of another user. To validate results of the label propagation algorithm, we apply a stratified cross (5-fold) validation to a set composed of 38,920 seed accounts. We train the algorithm using 80% of the seeds and we evaluate the performance on the remaining 20%. Finally, we compute precision and recall by reiterating the validation of the 5-folds. Both precision and recall scores show value around 0.89 with bounds from 0.88 to 0.90. Both the scores for liberals are about 0.87 with 0.85-0.88 bounds, while for conservatives the scores are around 0.93 with 0.92-0.93 bounds. To further validate the proposed approach, we use as a ground truth the political leaning of the media outlet that users shared in their profile, obtaining precision and recall scores in line with the previous approach.

[3]https://www.allsides.com/media-bias/media-bias-ratings
[4]https://mediabiasfactcheck.com/

Table 2: Users and tweets statistics

	Liberal	Conservative
Humans	386,391 (38.7%)	122,761 (12.3%)
Bots	82,118 (8.2%)	49,488 (4.9%)

(a) Number (percentage) of users per group

3.3 Human-Bot Interaction

We next introduce four metrics to estimate the effectiveness of bot actions in involving humans and, at the same time, measure to what extent humans rely upon, and interact with the content generated by social bots. Thereby, we propose the following metrics:

- *Retweet Pervasiveness* (*RTP*) measures the intrusiveness of bot-generated content in human-generated retweets:

$$RTP = \frac{\text{no. of human retweets from bot tweets}}{\text{no. of human retweets}} \quad (1)$$

- *Reply Rate* (*RR*) measures the percentage of replies given by humans to social bots:

$$RR = \frac{\text{no. of human replies to bot tweets}}{\text{no. of human replies}} \quad (2)$$

- *Human to Bot Rate* (*H2BR*) quantifies human interaction with bots over all the human activities in the social network:

$$H2BR = \frac{\text{no. of humans interaction with bots}}{\text{no. of humans activity}}, \quad (3)$$

  where the numerator counts for human replies/retweets to/of bots generated content, while the denominator is the sum of the number of human tweets, retweets, and replies.
- *Tweet Success Rate* (*TSR*) is the percentage of tweets generated by bots that obtained at least one retweet by a human:

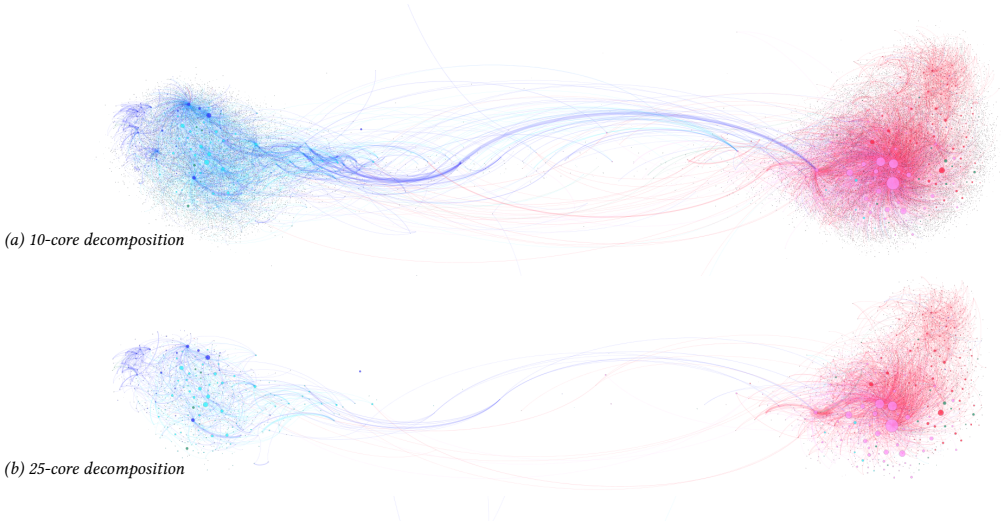
$$TSR = \frac{\text{no. of tweet retweeted at least once by a human}}{\text{no. of bots tweets}} \quad (4)$$

4 RESULTS

Next, we address the research questions discussed in the Introduction. We examine social bot partisanship and, accordingly, we analyze bots' strategies and measure the effectiveness of their actions in terms of human engagement.

4.1 RQ1: Bot Political Leaning

The combination of the outcome from the bot detection algorithm and the political ideology inference allowed us to identify four groups of users, namely Liberal Humans, Conservative Humans, Liberal Bots, and Conservative Bots. In Table 2a, we show the percentage of users per group. Note that percentages do not sum up to 100 as either the political ideology inference was not able to classify every user, or Botometer did not return a score, as we previously mentioned. In particular, we were able to assign a political leaning to 63% of bots and 67% of humans. We find that the liberal user population is almost three times larger than the conservative counterpart. This discrepancy is also present, but less evident, for the bot accounts, which exhibit an unbalance in favor of liberal bots. Further, we investigate the suspended accounts to inspect the consistency of this result. The inference algorithm attributed a political ideology to 63% of these accounts, which show once again the liberal advantage over the conservative faction (45% vs. 18%).



(a) 10-core decomposition

(b) 25-core decomposition

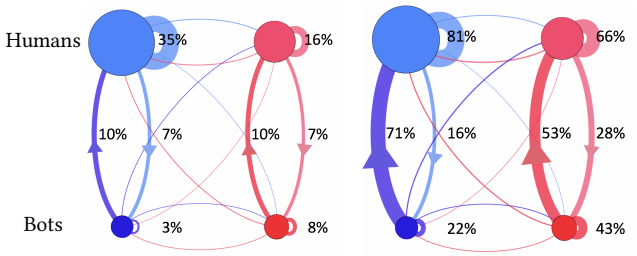
Figure 2: Political discussion over *(a)* the 10-core, and *(b)* the 25-core decomposition of the retweet network. Each node represents a user, while links represent retweets. Links with weight (i.e., frequency of occurrence) less than 2 are hidden to minimize visual clutter. Blue nodes represent liberal accounts, while red nodes indicate conservative users. Darker tones (blue and red) depict bots, while lighter tones (cyan and pink) relate to humans, and the few green nodes represent unclassified accounts. The link takes the same color of the source node (author of the retweet), whereas node size is proportional to the in-degree of the user.

Table 3: Top 20 hashtags generated by liberal and conservative bots. Hashtags in bold are not present in the top 50 hashtags used by the corresponding human group.

Liberal Bots	Conservative Bots
#MAGA	#BrowardCounty
#NovemberisComing	#MAGA
#TheResistance	#StopTheSteal
#GOTV	#WalkAway
#Florida	#WednesdayWisdom
#ImpeachTrump	#PalmBeachCounty
#Russia	#Florida
#VoteThemOut	#QAnon
#unhackthevote	#KAG
#FlipTheHouse	#IranRegime
#RegisterToVote	#Tehran
#Resist	#WWG1WGA
#ImpeachKavanaugh	**#Louisiana**
#GOP	#BayCounty
#MeToo	#AmericaFirst
#AMJoy	#DemocratsAreDangerous
#txlege	#StopTheCaravan
#FlipTheSenate	**#Blexit**
#CultureOfCorruption	**#VoteDemsOut**
#TrumpTrain	#VoterFraud

Figure 2 shows two k-core decomposition graphs of the retweet network. In a k-core, each node is connected with at least k other nodes. Figures 2a and 2b capture the 10-core and 25-core decomposition, respectively. Here, nodes represent Twitter users and link represent retweets among them. We indicate as *source* the user that retweeted the tweet of a *target* user. Colors represent the political ideology, with darker colors (red and blue) being bots and lighter colors (cyan and pink) being human users; size represents the in-degree. The graph is visualized using a force-directed layout [19], where nodes repulse each other, while edges attract their nodes. In our setting, this means that users are spatially distributed according to the amount of retweets between each other. The result is a network naturally split into two communities, where each side is almost entirely populated by users with the same political ideology. This polarization is also reflected by bots, which are embedded, with humans, in each political side. Two facts are worth noting: *(i)* as k increases, the left k-core appears to disrupt, while the right k-core remains well connected; and, *(ii)* as k increases, bots appear to outnumber humans, suggesting that bots may populate areas of the retweet network that are more central and better connected.

Next, we examine the topics discussed by social bots and compare them with the human counterparts. Table 3 shows the top 20 hashtags utilized by liberal and conservative bots. We highlight (in bold) the hashtags that are not present in the top 50 hashtags used by the corresponding human group to point out the similarities and differences among the groups. In this table, we do not take into account hashtags related to the keywords used in the data collection (such as #elections, #midterms), and hashtags used to support the political group (such as #democrats, #liberals, #VoteRed(or Blue)ToSaveAmerica) as (i) the overlap between bot and human hashtags is noticeable when these terms are considered (in the interest of space, we do not show this result in Table 3), and (ii) we aim to narrow the analysis to specific topics and inflammatory content, inspired by [27]. Moreover, we used an enlarged subset of hashtags for the human groups to further strengthen the differences and, at the same time, to better understand the objective of social bots. Although bots and humans share the majority of hashtags, two main differences can be noticed. First, conservative bots abide by the corresponding human counterpart more than the liberal bots. Second, liberal bots focus on more inflammatory and



(*a*) Overall interactions (*b*) Group-based interactions

Figure 3: Interactions according to political ideology

provocative content (e.g., #ImpeachTrump, #ImpeachKavanaugh, #FlipTheSenate) w.r.t. conservative bots.

4.2 RQ2: Bot Activity and Strategies

In this Section, we investigate social bot activity based on their political leaning. We explore their strategies in interacting with humans and the degree of embeddedness in the social network.

Table 2b depicts the number (and percentage) of tweets generated by each group. Despite the group composed of conservative bots is the smallest in terms of number of accounts, it produced more tweets than liberal bots and closely approaches the number of tweets generated by the human counterpart. The resulting *tweet per user ratio* shows that conservative bots produce 7.4 tweets per account, which is more than twice the ratio related to the liberal bots (3.5), almost the double of the human counterpart (3.9), and nearly three times the ratio of liberal humans (2.5).

To investigate the interplay between bots and humans, we consider the previously described retweet network. Figure 3 shows the interaction among the four groups. We maintain the same color mapping described before, with darker color (on the bottom) representing bots and lighter color (on top) indicating humans. Node size is proportional to the percentage of accounts in each group, while edge size is proportional to the percentage of interactions between each group. In Figure 3a, this percentage is computed considering all the interactions in the retweet network, while in Figure 3b we consider each group separately, therefore, the edge size gives a measure of the group propensity to interact with the other groups. Consistently with Figure 2, we observe that there is a limited amount of interaction between the two political sides. The majority of interactions are either intra-group or between groups of the same political leaning. From Figure 3b, we can observe that the two bot factions adopted different strategies. Conservative bots balanced their interactions by retweeting group members 43% of the time, and the human counterpart 52% of the time. On the other hand, liberal bots mainly retweeted liberal humans (71% of the time) and limited the intra-group interactions to the 22% of their retweet activity. Interestingly, conservative humans interacted with the conservative bots (28% of the time) much more than the liberal counterpart (16%) with the liberal bots. To better understand these results and to measure the extent of human engagement with bots, in the next Section we evaluate the four metrics introduced earlier in this paper.

Table 4: Average network centrality measures

	Liberal	Conservative		Liberal	Conservative
Humans	$2.66 \cdot 10^{-6}$	$4.14 \cdot 10^{-6}$	Humans	$2.52 \cdot 10^{-6}$	$4.24 \cdot 10^{-6}$
Bots	$3.70 \cdot 10^{-6}$	$7.81 \cdot 10^{-6}$	Bots	$2.53 \cdot 10^{-6}$	$6.22 \cdot 10^{-6}$

(a) Out-degree centrality **(b) In-degree centrality**

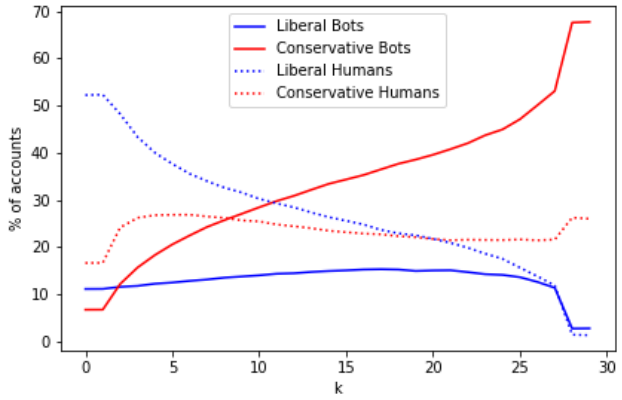


Figure 4: k-core decomposition, liberal vs. conservative users

Finally, we examine the degree of embeddedness of both humans and bots within the retweet network. For this purpose, we first compute different network centrality measures, and then we adopt the k-core decomposition technique to identify the most central nodes in the graph. In Table 4, we show the average out- and in-degree centrality for each group of users. Out-degree centrality measures the quantity of outgoing links, while in-degree centrality considers the number of of incoming links. Both of these measures are normalized by the maximum possible degree of the graph. Overall, conservative groups have higher centrality measures than the liberal ones. We can notice that conservative bots achieve the highest values both for the out- and in-degree centrality. To further investigate bots embeddedness in the social network, we use the k-core decomposition. The objective of this technique is to determine the set of nodes deeply embedded in a graph. The k-core is a subgraph of the original graph in which every node has a degree equal to or greater than a given value k. We extracted the k-cores from the retweet network by varying k in the range between 0 and 30. Figure 4 depicts the percentage of liberal and conservative users as a function of k. We can notice that, as k grows, the fraction of conservative bots increases, while the percentage of liberal bots remains almost stationary. On the human side, the liberal fraction drops with k, whereas the conservative percentage remains approximately steady. Overall, conservative bots sit in a more central position in the social network and are more deeply connected if compared to the liberal counterpart.

4.3 RQ3: Bot Effectiveness in Human Engagement

In this Section, we aim to estimate the effectiveness of bot strategies in involving humans and measure to what extent humans rely upon, and interact with the content generated by social bots. We examine the effect of bot activities by means of the four metrics described in

Table 5: Bot Effectiveness in Human Engagement

Metric	Liberal Bots	Conservative Bots
RTP	14.1%	25.6%
RR	4.5%	15.5%
$H2BR$	12.3%	23.2%
TSR	35.3%	35.0%

Section *Human-Bot Interaction*. We evaluate each political side separately, thus, we compare the interaction between bots and humans with the same leaning. In Table 5, we depict the results for each group of bots. Diverse aspects are worthy of consideration. We can observe that conservative bots are significantly more effective in involving humans in their conversations than the liberal counterpart. Although the TSRs of the *red* and *blue* bots are comparable, the gap between the two groups, with respect to the other metrics, is significant. To carefully interpret this result, it should also be noticed that (i) the TSR is inversely proportional to the number of tweets generated by bots, and (ii) conservative bots tweeted more than the liberal counterpart, as depicted in Table 2b. Overall, conservative bots received a larger degree of interaction with (and likely trust from) human users. In fact, conservative humans interacted with the bot counterpart almost twice with retweets (RTP), and more than three times with replies (RR) if compared to the liberal group. Finally, the $H2BR$ highlights a remarkable amount of human activities that involve social bots: almost one in four actions performed by conservative humans goes towards red bots.

5 CONCLUSIONS & FUTURE WORK

In this work, we conducted an investigation to analyze social bots activity during the 2018 US Midterm election. We showed that social bots are embedded in each political wing and behave accordingly. We observed different strategies between conservative and liberal bots. Specifically, conservative bots stand in a more central position in the social network and abide by the topic discussed by the human counterpart more than the liberal bots, which in turn exhibit an inflammatory attitude. Further, conservative bots balanced their interaction with humans and bots of the red wing, whereas liberal bots focused mainly on the interplay with the human counterpart.

Finally, we inspected the effectiveness of these strategies in terms of human engagement and recognized the strategy of the conservative bots as the most effective. However, these results open the door to further interpretation and discussion. Are conservative bots more effective because of their strategy or because of the human ineptitude to distinguish their nature? This, and related analysis, will be expanded in future work.

Acknowledgements. The authors gratefully acknowledge support by the Air Force Office of Scientific Research (award #FA9550-17-1-0327). L. Luceri is funded by the Swiss National Science Foundation (SNSF) via the CHIST-ERA project *UPRISE-IoT*.

REFERENCES

[1] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *Int. Conference on Advances in Social Networks Analysis and Mining*. 258–265.
[2] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who Falls for Online Political Manipulation? *Companion of the The Web Conference 2019* (2019).
[3] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).
[4] Olga Boichak, Sam Jackson, Jeff Hemsley, and Sikana Tanupabrungsun. 2018. Automated Diffusion? Bots and Their Influence During the 2016 US Presidential Election. In *International Conference on Information*. Springer, 17–26.
[5] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.
[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[7] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. DeBot: Twitter Bot Detection via Warped Correlation.. In *ICDM*. 817–822.
[8] Zhouhan Chen and Devika Subramanian. 2018. An Unsupervised Approach to Detect Spam Campaigns that Use Botnets on Twitter. *arXiv:1804.05232* (2018).
[9] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*.
[10] Ashok Deb, Luca Luceri, Adam Badawy, and Emilio Ferrara. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. *Companion of the The Web Conference 2019* (2019).
[11] Emilio Ferrara. 2015. Manipulation and abuse on social media. *ACM SIGWEB Newsletter* Spring (2015), 4.
[12] Emilio Ferrara. 2017. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday* 22, 8 (2017).
[13] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.
[14] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. 2016. Detection of promoted social media campaigns. In *ICWSM*. 563–566.
[15] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378.
[16] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5, 1 (2019), eaau4586.
[17] Philip N Howard, Gillian Bolsover, Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. 2017. Junk news and bots during the US election: What were Michigan voters sharing over Twitter. *CompProp, OII, Data Memo* (2017).
[18] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv:1901.11162* (2019).
[19] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9, 6 (2014), e98679.
[20] Sneha Kudugunta and Emilio Ferrara. 2018. Deep Neural Networks for Bot Detection. *Information Sciences* 467, October (2018), 312–322.
[21] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. 2017. Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots. *Plos One* 12, 9 (2017), e0184148.
[22] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.
[23] Iacopo Pozzana and Emilio Ferrara. 2018. Measuring bot and human behavioral dynamics. *arXiv:1802.04286* (2018).
[24] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *PNAS* (2019), 201805871.
[25] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 4787.
[26] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[27] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115, 49 (2018), 12435–12440.
[28] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, et al. 2016. The DARPA Twitter Bot Challenge. *Computer* 49, 6 (2016).
[29] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Int. AAAI Conference on Web and Social Media*. 280–289.
[30] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[31] Samuel Woolley and Douglas Guilbeault. 2017. Computational propaganda in the United States of America: Manufacturing consensus online. *CompProp* (2017).
[32] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* (2019), e115.

A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections.

Caio Machado
Oxford Internet Institute
Oxford University, Oxford, United Kingdom,
caio.machado@oii.ox.ac.uk

Beatriz Kira
Department of Law
University of Sao Paulo, Sao Paulo, Brazil,
beatrizkira@gmail.com

Vidya Narayanan
Oxford Internet Institute
Oxford University, Oxford, United Kingdom,
vidya.narayanan@oii.ox.ac.uk

Bence Kollanyi
Oxford Internet Institute
Oxford University, Oxford, United Kingdom, bence.kollanyi@oii.ox.ac.uk

Philip N. Howard
Oxford Internet Institute
Oxford University, Oxford, United Kingdom, philip.howard@oii.ox.ac.uk

ABSTRACT

There are rising concerns over the spread of misinformation in WhatsApp groups and the potential impact on political polarization, hindrance of public debate and fostering acts of political violence. As social media use becomes increasingly widespread, it becomes imperative to study how these platforms can be used to as a tool to spread propaganda and manipulate audience groups ahead of important political events. In this paper, we present a grounded typology to classify links to news sources into different categories including 'junk' news sources that deliberately publish or aggregate misleading, deceptive or incorrect information packaged as real news about politics, economics or culture obtained from public WhatsApp groups. Further, we examine a sample of 200 videos and images, extracted from a sample of WhatsApp groups and develop a new typology to classify this media content. For our analysis, we have used data from 130 public WhatsApp groups in the period leading up to the two rounds of the 2018 Brazilian presidential elections.

CCS CONCEPTS

• Social media networks • Online social networks

KEYWORDS

WhatsApp, Misinformation, Brazilian Elections, Typology

ACM Reference format:

Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi and Philip N. Howard. 2019. A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections.. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316738

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316738

1 Introduction

WhatsApp, the private messaging application, has emerged as a popular tool for inter-personal communication in several countries around the globe. The role of social media in spreading political misinformation has been well-documented in the literature [1] [2]. However, our previous research noted that information sharing and political conversations have progressively moved away from public platforms like Twitter and into more private discussion spaces like WhatsApp or Facebook Messenger, particularly in countries like Mexico and Brazil [3] [4]. Consequently, this paper presents our preliminary study of the role of WhatsApp in political communication and propaganda, using data from Brazilian WhatsApp groups during the period leading up to the two rounds of the Brazilian presidential elections. This work seeks to answer the following questions: (1) What types of news sources are being shared within the WhatsApp groups studied? (2) What types of content are being shared? (3) How much of the media content --- image and video files shared can be considered to be polarizing content?

2 Sampling and Methods

In order to examine how political news and information was being circulated within WhatsApp groups, we collected data related to the Brazilian presidential elections, given the enormous popularity of the messaging app in the country and amid reports that it was being used widely as a campaigning tool. We joined political WhatsApp groups dedicated to discussion of politics, news and current affairs with publicly available invite links. These groups were easily accessible and were preferred channels to disseminate political information. Our approach relied on information available on the web, indexed directly by search engines or available in public repositories that can be found through search engines. Using

Google Search and DuckDuckGo we obtained access to a few repositories, two of which were particularly active, namely Gruposwhats.com and Gruposdezap.com. We also joined WhatsApp groups made publicly available on campaign pages, but many links were quickly revoked, possibly due to attacks by supporters from other candidates. In these repositories, we used keywords that would identify parties, presidential candidates or phrases used by the main campaigns. General keywords such as "Politic*", "Right", "Left", "Liberal" and "Conservative" (in Brazilian Portuguese) were used as search queries for both the search engine and the repositories search tool, in order to identify political groups across the spectrum of Brazilian politics. Some of the main repositories used in this study were: http://www.gruposwhats.com/, https://www.gruposdozapzap.com.br/, https://www.linksdegruposzap.com/. Public links made available in repositories has a high turnover and were often revoked. On the other hand, by joining invite links made available within the WhatsApp groups it was possible to join a substantial number of political groups. Using this snowball sampling method, we joined 130 public WhatsApp groups. Groups varied in size and had a high turnover of participants that joined and left groups frequently. Many of our groups reached the 256-user limit. According to our ethics protocol, sought the consent of group members in the groups we joined from the repositories and then snowballed into other groups via public links and invitations circulating in those chats. To seek consent, we posted a message at the time of joining the group announcing ourselves as researchers and requesting group members to contact us if they had any issues with their data being used for this study. Further, all data for this analysis was strictly anonymised and only included media links, images and audio-visual content shared on this thread. We used no personal data or messages to inform our research findings. Between September 21 – October 20 we collected 298,892 WhatsApp messages and received, 99,988 media files. We note that this snowball sampling strategy does not ensure that all existing public WhatsApp groups, are adequately represented. The closed nature of the platform means that it is very difficult to comprehensively map the network of public groups in a region. Also, this data collection method may introduce a bias which is linked to our initial dataset, where groups might share invites to other groups with similar political alignments.

We accessed the chat history from 130 public groups and extracted 50,795 URLs shared within these groups. Out of these 5,723 URLs pointed to the whatsapp.com domain itself. After examining these links and finding that they were all WhatsApp group invites, we removed the links from our dataset. We analyzed the remaining 45,072 links and successfully labeled 38,800 links, achieving 86.1% coverage of the dataset. This dataset, apart from containing links to diverse news sources, also includes links to other platforms such as YouTube, Facebook and Twitter. Due to the large number of YouTube and Facebook links, we labeled all unique links from these two platforms that were shared five times or more on WhatsApp. We treated every post from the same public Facebook public page as content coming from one source. Similarly, in the case of YouTube, we used the same label for all videos originating from the same channel. All other URLs were classified on a domain or sub-domain level. The coding was done according to a rigorous and iterative coding process and by using a previously developed grounded typology. This typology was tested and refined by analyzing links shared on social media platforms, including Twitter and Facebook, during elections and other high salience political events across multiple countries [5]. The links were coded by a team of coders with high inter-coder reliability measured by the Krippendorff's alpha which was 0.84. The existing literature on coder reliability metrics concludes that this provides a high level of reliability [6]. The typology explaining our content classification is detailed below:

Professional News Content

- Major News Brands. This is political news and information by major newspapers, broadcasting or radio outlets, as well as news agencies.
- Local News. This content comes from local and regional newspapers, broadcasting and radio outlets, or local affiliates of major news brands.
- New Media and Start-ups. This content comes from new media and digitally native publishers, news brands and start-ups.
- Tabloids. This news reporting focuses on sex, crime, astrology and celebrities, and includes yellow press publications.

Professional Political Content

- Government. These links are to websites of branches of government or public agencies.
- Experts. This content takes the form of white papers, policy papers or scholarship from researchers based at universities, think tanks or other research organizations.
- Political Party or Candidate. These links are to official content produced by a political party or candidate campaign, as well as the parties' political committees.

Polarizing and Conspiracy Content

- Junk News and Information. These sources deliberately publish misleading, deceptive or incorrect information purporting to be real news about politics, economics or culture. This content includes various forms of propaganda and ideologically extreme, hyper-partisan or conspiratorial news and information. To be classified as Junk News and Information, the source must fulfill at least three of these five criteria:
- *Professionalism*: These outlets do not employ standards and best practices of professional journalism. They refrain from providing clear information about real authors, editors, publishers and owners. They lack transparency and accountability, and do not publish corrections on debunked information.

- *Style*: These outlets use emotionally driven language with emotive expressions, hyperbole, ad hominem attacks, misleading headlines, excessive capitalization, unsafe generalizations and logical fallacies, moving images, and lots of pictures and mobilizing memes.
- *Credibility*: These outlets rely on false information and conspiracy theories, which they often employ strategically. They report without consulting multiple sources and do not fact-check. Sources are often untrustworthy and standards of production lack reliability.
- *Bias*: Reporting in these outlets is highly biased, ideologically skewed or hyper-partisan, and news reporting frequently includes strongly opinionated commentary.
- *Counterfeit*: These sources mimic established news reporting. They counterfeit fonts, branding and stylistic content strategies. Commentary and junk content that is stylistically disguised as news, with references to news agencies and credible sources, and headlines written in a news tone with date, time and location stamps.
- Russia. This content is produced by known Russian sources of political news and information.
- Other Political News and Information
- Political Commentary Blogs. Political blogs that employ standards of professional content production such as copy- editing, as well as employ writers and editorial staff. These blogs typically focus on news commentary rather than neutral news reporting on a news cycle and are often opinionated or partisan.
- Citizen, Civil Society and Civic Content. These are links to content produced by independent citizen, civic groups, civil society organizations, watchdog organizations, fact- checkers, interest groups and lobby groups representing specific political interests or agendas. This includes blogs and websites dedicated to citizen journalism, personal activism, and other forms of civic expression that display originality and creation that goes beyond curation or aggregation. This category includes Medium, Blogger and WordPress, unless a specific source hosted on either of these pages can be identified."

Further, we analyzed a random sample of 200 images and 200 videos that were collected between 21 September - 28 September, from our data sample. We initially classified videos and images into 5 main parent categories broadly on the basis of their political affiliations:

- Pro-PSL: in favor of the Social Liberal Party
- Anti-PSL: against the Social Liberal Party
- Pro-PT: in favor of the Worker's Party
- Anti-PT: against the Worker's Party
- Other Parties

These groups are representative of the political forces dominating the public debate in the Brazilian elections, particularly in the run up to the second round. Under each one of the parent categories, we further classified the media content under the following sub-categories:

- Campaign and Support Material – Broadly supportive of the candidate, using verifiable facts and figures, pertaining to social, economic and political issues.
- Polarizing/Junk Content –
- Style: Content uses emotive expressions, hyperbole, misleading information, unsafe generalizations and logical fallacies.
- Credibility: Content promotes false information and conspiracy theories, which is often employed strategically.
- Bias: Content is highly biased, ideologically skewed or hyper-partisan. Promotes inflammatory viewpoints.
- Discrediting tactics: Content discrediting opponents, by indulging in smear campaigns, ad hominem attacks, falsehoods about personal lives, attacks using religious themes/iconography.
- Religion – Pertaining to religious practices, references to religious texts and values, claiming alignment with core religious values from political leaders, claiming support of prominent religious leaders for the campaign.
- Celebrities – Prominent social figures in the arts, TV stars, sports personalities, senior officials in law enforcement and other government institutions voicing strong support for candidates.
- Hate, Gore and Porn – Hate speech/content directed against individuals, minority groups like women, LQBT community, against ideologies, as well as videos depicting acts of extreme violence or obscenities.
- Satire – Content that uses humor, irony or exaggeration to comment on political and social issues.
- Other – Lifestyle, goodwill messages and content that doesn't fit the above categories.

3 Findings

First, we analyzed the types of news content shared between 21 September and 20 October. Looking specifically at links being shared within WhatsApp groups, we note that junk news sources account for 13.1% of all links in our dataset (see Table 1). On the other hand, the relatively low number of professional political sources 2.7%, could be caused by one or both of two things: first, the adaptation to the cultural use of WhatsApp, which is an interpersonal platform and tends to simulate a more private environment; second, the strategic use of WhatsApp to disseminate junk content, benefitting from reduced control mechanisms and debilitated law enforcement mechanisms. Though there have been attempts to implement technical

solutions on WhatsApp to restrict the viral spreading of junk news, there are still no means of monitoring or tracing content, nor of enabling the removal of illegal content.

Another interesting observation is the very high number of YouTube links in our WhatsApp dataset. The 17,702 links, pointing to YouTube videos account for close to 40% of the shared links, suggesting a strong interaction across these two platforms, where affordances and sharing strategies differ. We labeled 11,255 links out of the total number of YouTube links from our dataset, using our grounded typology and the remaining 6,447 YouTube links were classified as Video/Image sharing (see Tables 1 & 2). Further links to YouTube and Facebook are used to expose users to high percentages of junk content, namely 30.9% on YouTube and 42.3% on Facebook (see Table 2). Moreover, the general culture around WhatsApp use in Brazil favors a very open and participative relationship on social media even in public groups, sharing entertainment, gossip and sometimes violence [7]. This behavior is a trend even on Facebook, where users tend to be very active and public about their activities and posts. Following our overall data analysis, we analyzed a random sample of media content extracted between 21 September - 28 September 2018. The purpose of this collection was to provide a broad overview of the media content being shared in public WhatsApp groups during the election campaign for the first round. Our analysis is presented in Tables 3 & 4. Looking at the different types of content, we note that videos are an important resource for the dissemination of polarizing and conspiratorial content. They are particularly effective because they are easily edited and can convey information using loaded and appealing imagery. Videos can simulate the production of content from users as "eye witnesses" of certain events or simulate personal statements of citizens and specialists, misleading the users. We coded over two thirds of the YouTube links shared within the WhatsApp groups. The approach excluded the links that were shared less than 5 times, which accounts for the long-tail distribution of videos which were in the last third of the total links shared. This means that the sample of YouTube links coded is highly representative of the volume of links being disseminated, even though it isn't representative with regards to the variety of content. These diverse, but less frequently shared links could point towards to non-political user activity on the platform such as sharing entertainment videos. A similar trend was observed in Facebook. Thus, although WhatsApp has its own affordances for sharing and disseminating media within the platform, bringing users to YouTube could be a strategy for exposing them to a network of content based on YouTube recommendations.

While authorities and researchers focus on platforms in an isolated manner, it is hard to evaluate how the exposure to one link on one platform could lead to contact with polarizing content when the user follows the path of related content. For example, while WhatsApp's media sharing was under scrutiny of authorities, bringing users to YouTube could be a strategy to avoid detection. The collection of images and video files from the WhatsApp conversations had exploratory purposes. Rather than aiming at representativeness of all the media being disseminated through WhatsApp, this analysis reflects important findings of our research that can offer directions for future research. In this regard, we coded images and videos according to a high-level typology, which would allow us to broadly classify the political orientation of content circulating and identify patterns of polarizing and conspiratorial material. For the purposes of this research we combined categories that supported a specific candidate or that opposed his adversaries under a single group. For example, we combined content that specifically attacked Jair Bolsonaro (Anti-PSL) and content that supported Fernando Haddad (Pro-PT) into a single unified category (Anti-PSL / Pro-PT), in order to better depict the polarized dispute between the candidates. As seen in Tables 3 & 4, a majority of the videos and images are classified as Anti-PT/Pro-PSL. We believe that this is due to: (1) the use of WhatsApp as a broadcast system was an innovative strategy and only one political party had developed this communication structure. (2) further specific foreign advertising agencies were hired to offer this service, which was something other political parties had not yet caught onto [8]. This methodological choice is justified because the online debate was dominated by these two figures and virtually all of the political content was used either to support and/or attack one of these two candidates. This approach is supported by the fact that we counted only 14 videos and 22 images supporting other candidates. As we can see from the Tables 3 & 4, the number of Polarizing and Junk images and videos is high, and it is worth pointing out that the categories Satire and Hate, Gore and Porn had also had high numbers. The media we collected has been very similar across the groups supportive of the same candidate. The presence of these elements points towards the intense use of visual and emotional strategies to influence the user. Media files (99,988) accounted for 33% of all messages collected (298,892) that included links to news sources, text messages and media files. Therefore, a significant proportion of overall content being disseminated are images, audio and video files, illustrating the prolific use of audio-visual strategies to disseminate news and information in WhatsApp groups.

Though the data analyzed is insufficient for drawing conclusions about the comparative use of videos and images between the candidates, it shows that the media affordances of WhatsApp are being intensely exploited for the dissemination of political content. This is an important finding, because images and videos don't fit the concept of "news", in the sense of being reported by periodicals or media outlets. Such media rarely have the intention of simulating the authority of media sources, as is common with junk news being shared online. For these purposes, images and videos operate with a slightly different rhetoric than the one established by the "fake news" debate, since they usually don't point to any source or attempt to pass any credit or authority. They are strategically constructed rumors, that will distort and attack proposals from the opposing party or that will evoke emotional reactions from the users. More importantly, Bolsonaro's campaign has repeatedly attacked major media outlets as "fake news" outlets and defended WhatsApp as a free space where users can report 'first-hand' experiences and bypass the biases of the traditional media

gatekeepers [9]. Within this context, the rhetoric of memes and videos is able to thrive without indicating any sort of source, since they appeal to users as a truth that was formerly concealed under the traditional media monopoly. This also fosters speculative behavior from the users, that might forward content simply for supporting their views or for raising awareness. Reports have indicated that out of the top 50 most widely diffused images, 46 were fake or distorted [10]. These findings are important indicators of the strategic use of images to disseminate polarizing content, which spread widely under the veil of satires, memes, or openly as hoaxes, and whose source is almost impossible to verify. Such content appears to be user produced, especially when disseminated on WhatsApp, where the content is forwarded by users and hosted on the users' phones. Shifting from a direct public broadcasting of information (ex. Facebook, Twitter) to a broadcast strategy via a network of groups can be highly effective in reaching out to users in a medium that is perceived as more intimate, while making it extremely difficult to trace back to the disseminating accounts or even to filter content distribution.

4 CONCLUSIONS AND FUTURE WORK

This research represents a preliminary study of the use of WhatsApp as a tool for political propaganda in Brazil. From analyzing links to news sources and random sample of images and videos we found that: (1) In Brazil, WhatsApp presents an extremely low number of professional political content and a high number of junk news content; (2) Information spreading on WhatsApp relies intensely on the dissemination media files, which don't use the same rhetoric as junk news sources, not attempting to simulate authority to credit information; (3) Content dissemination strategies within WhatsApp groups often resort to hate speech and deception to achieve viral dissemination. Our investigations indicate that visual metaphors are being heavily used within WhatsApp groups to distort information and manipulate users. The analysis of this content at scale, will require a combination of both quantitative and qualitative methods. In order to analyze large volumes of visual content, we propose in our future work to explore automated image recognition techniques to develop broad clustering strategies, for instance identify the most popular images that have been shared within the groups in our data sample, and then apply a detailed typology for qualitatively classifying clustered media files into various categories. We believe that misinformation campaigns in the future, specifically in messaging apps like WhatsApp will increasingly use visual media to spread propaganda and carry out influence operations. Therefore, the work of analyzing such content will be vital to ensuring that such content does not endanger democratic processes by polarizing societies and deepening fault lines within communities.

Table 1: Types of News and Information Shares on WhatsApp

Type of Source	N	%
Professional News Outlets		
News Brands	10,503	27.1
Tabloids	70	0.2
Subtotal	10,574	27.3
Professional Political Content		
Political Party or Candidate	625	1.6
Government	287	0.7
Expert	131	0.3
Subtotal	1,043	2.7
Polarizing & Conspiratorial Content		
Junk News	5,093	13.1
Subtotal	5,093	13.1
Other Political News & Information		
Video/Image Sharing	6,524	16.8
Citizen or Civil Society	1,828	4.7
Political Commentary Blog	1,724	4.4
Portals, Search & Aggregators	909	2.3
Fundraising & petitions	476	1.2
Remaining categories**	436	1.1
Subtotal	11,897	30.7
Others		
Social Media Platforms*	3,086	8.0
Portals, Search & Aggregators	4,067	10.5
Fundraising & petitions	1,303	3.4
Remaining categories***	1,737	4.5
Subtotal	10,193	26.3
Total	38,800	100

Source: Authors' calculations from data collected from Sept 21 to Oct 10, 2018.

* Besides other image and video sharing platforms, YouTube links were labelled as SHC unless they were coded separately. Similarly, Facebook links were labelled as SM unless coded individually (For more details, see Table 2).

** Political Humor, Lifestyle, Religion, Cloud Services, and Other Political were collapsed as these categories constituted low percentage.

*** Not Available and Link Shorteners were collapsed as these categories constituted low percentage.

Table 2: Links Shared to Different Platforms on WhatsApp.

Source of the Link	Total Links (N)	Links Coded (N)	Junk Links (N)	Junk Links (%)
YouTube Video	17,702	11,255	3,477	30.9
Facebook Post and Page	1,549	494	209	42.3
Other news and Information Sources	25,822	19,549	1,407	7.2

Source: Authors' calculations from data collected from Sept 21 to Oct 10 2018.

Table 3: Analysis of a Random Sample of 200 Videos.

Categories	Anti-PT/ Pro-PSL	Anti-PSL / Pro-PT	Other Political	Non-Political
Campaign and Support Material	56	4	6	0
Polarizing / Junk Content	35	0	0	1
Religion	4	1	0	3
Celebrities	2	0	0	0
Hate, Gore and Porn	12	0	0	7
Satire	16	1	2	12
Lifestyle	0	0	0	3
Remaining Categories	6	5	6	18

Source: Authors' calculations from data collected from Sept 21 to Sept 28, 2018.

Table 4: Analysis of a Random Sample of 200 Images.

Categories	Anti-PT/ Pro-PSL	Anti-PSL / Pro-PT	Other Political	Non-Political
Campaign and Support Material	37	9	16	0
Polarizing / Junk Content	33	4	1	0
Religion	0	1	0	5
Celebrities	7	0	0	0
Hate, Gore and Porn	4	1	0	2
Satire	33	2	2	5
Lifestyle	0	0	0	13
Remaining Categories	3	0	3	19

Source: Authors' calculations from data collected from Sept 21 to Sept 28, 2018.

5 ETHICAL CONSIDERATIONS

We adhered to strict professional and ethical guidelines. This includes safe data storage policies, concerted attempts to include a diverse range of participants, and open explanation of project methods and goals with participants. Our researchers agreed not to interact in conversation or collect data on specific users of messages. Chat logs were accessed, downloaded, and pseudonymized before our research staff could access them for analysis. We only undertook quantitative data analysis of these chat logs, extracted links to news sources, classified other content like memes and jokes, and not use personal quotes or opinions sourced from these transcripts, in our research.

5 ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the European Research Council, "Computational Propaganda: Investigating the Impact of Algorithms and Bots on Political Discourse in Europe," Proposal 648311, 2015-2020, Philip N. Howard, Principal Investigator. Project activities were approved by the University of Oxford's Research Ethics Committee, CUREC OII C1A 15-044. For supporting our Election Observatory and our research in Europe we are grateful to Hewlett Foundation and Omidyar Network. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Oxford or our funders.

REFERENCES

[1] V. Narayanan, V. Barash, J. Kelly, B. Kollanyi, L.-M. Neudert, and P. N. Howard, "Polarization, Partisanship and Junk News Consumption over Social Media in the US," Oxford Internet Institute, Oxford, Data Memo, Feb. 2018.

[2] P. N. Howard, S. Woolley, and R. Calo, "Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration," *Journal of Information Technology & Politics*, pp. 1–13, Apr. 2018.

[3] M. Glowacki *et al.*, "News and Political Information Consumption in Mexico: Mapping the 2018 Mexican Presidential Election on Twitter and Facebook," Data Memo 2018.2, 2018.

[4] C. Machado *et al.*, "News and Political Information Consumption in Brazil: Mapping the First Round of the 2018 Brazilian Presidential Election on Twitter." Computational Propaganda Project, 05-Oct-2018.

[5] P. N. Howard, B. Kollanyi, G. Bolsover, S. Bradshaw, and L.-M. Neudert, "Junk News and Bots during the U.S. Election: What Were Michigan Voters Sharing Over Twitter?," Mar. 2017.

[6] K. Krippendorff, "Reliability in content analysis," *Human communication research*, vol. 30, no. 3, pp. 411–433, 2004.

[7] J. Spyer, "Social Media in Emergent Brazil," *UCL Press*. [Online]. Available: https://www.ucl.ac.uk/ucl-press/browse-books/social-media-in-emergent-brazil. [Accessed: 04-Nov-2018].

[8] I. T. S. Rio, "Computational Power: Automated Use of WhatsApp in the Elections," *ITS FEED*, 26-Oct-2018. [Online]. Available: https://feed.itsrio.org/computational-power-automated-use-of-whatsapp-in-the-elections-59f62b857033. [Accessed: 02-Mar-2019].

[9] "Esse jornal se acabou, diz Bolsonaro ao Jornal Nacional sobre a Folha," *Folha de S.Paulo*, 29-Oct-2018. [Online]. Available: https://www1.folha.uol.com.br/poder/2018/10/esse-jornal-se-acabou-diz-bolsonaro-ao-jornal-nacional-sobre-a-folha.shtml. [Accessed: 04-Nov-2018].

[10] "Como funciona a máquina de WhatsApp que pode eleger Bolsonaro - Época." [Online]. Available: https://epoca.globo.com/como-funciona-maquina-de-whatsapp-que-pode-eleger-bolsonaro-23180627. [Accessed: 04-Nov-2018].

Institutional Counter-disinformation Strategies in a Networked Democracy

Jonathan Stray
Columbia Journalism School
New York, NY, USA, jms2361@columbia.edu

ABSTRACT

How should an organized response to disinformation proceed in a 21st century democratic society? At the highest level, what strategies are available? This paper attempts to answer these questions by looking at what three contemporary counter-disinformation organizations are actually doing, then analyzing their tactics. The EU East StratCom Task Force is a contemporary government counter-propaganda agency. Facebook has made numerous changes to its operations to try to combat disinformation, and is a good example of what platforms can do. The Chinese information regime is a marvel of networked information control, and provokes questions about what a democracy should and should not do. The tactics used by these organizations can be grouped into six high level strategies: refutation, exposure of inauthenticity, alternative narratives, algorithmic filter manipulation, speech laws, and censorship. I discuss the effectiveness and political legitimacy of these approaches when used within a democracy with an open Internet and a free press.

CCS CONCEPTS

• **Security and privacy~Social aspects of security and privacy** • **Social and professional topics~Censorship** • Information systems~Content ranking

KEYWORDS

Disinformation, policy, democracy, refutation, content ranking, censorship

ACM Reference format:

Jonathan Stray. 2019. Institutional Counter-disinformation Strategies in a Networked Democracy. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316740

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316740

1 Institutional Responses to Disinformation

There has been a great deal written about responses to online disinformation within the computer science, social science, and political science communities in the last few years. Such work includes technical approaches such as automated fact checking , psychological experiments concerning belief and persuasion , documentation of active disinformation campaigns , and proposed responses to specific threats .

There is also a need for a big-picture view. A robust societal response to disinformation will require long-lived institutions executing politically acceptable strategies. Institutionalized disinformation response within democracies peaked during the Cold War. At that time the U.S. had multiple agencies and programs tasked with countering disinformation and getting out an American narrative , the most significant of which was the United States Information Agency (1953-1999) which once had a staff of thousands and a budget of billions .

While there are good descriptions of Cold War counter-disinformation institutions, there is much less documentation of current organized efforts which must deal with the networked world. There are a few theory-driven analyses (such as Hellman and Wagnsson's typology of strategies), but little that synthesizes current practice. This paper attempts to fill this gap by outlining the work of three contemporary counter-disinformation organizations, then categorizing their tactics into higher-level strategies.

Having identified these strategies, this paper considers which are effective and legitimate in a pluralistic democracy. However, I make no attempt to define "disinformation." Adjudicating which speech is harmful is a profound problem with millennia of history. Instead, the emphasis here is on potential responses and their alignment with the values of an open society, *given that* a particular narrative is held to be damaging and worthy of countering.

Section 2 relates brief case studies of three organizations. Section 3 groups and discusses the types of tactics used across these organizations. Each strategy has certain advantages and disadvantages in terms of efficacy (how well it works) and legitimacy (alignment with the values of an open society.) Section 4 concludes.

2 Institutional Case Studies

2.1 Case study: the East StratCom Task Force

The EU East StratCom Task Force was created in 2015 at the direction of the European Council to respond to "Russia's ongoing disinformation campaigns". It is a modern example of a government counter-propaganda organization, and focusses on tracking and rebutting state propaganda.

It currently has just 14 full time staff and funding of €1.1 million per year . This small team of coordinates with a volunteer network of some 500 journalists, diplomats, NGO staff, and interested amateurs who assist in media monitoring and analysis in many countries . The unit publishes under the "EU vs. Disinfo" brand, and produces the weekly "Disinformation Review" along with rebuttals of specific items of disinformation. The content is majority English with substantial Russian.

The unit reports that it debunked 1310 "disinformation cases" in 2017 and over 4,700 since founding . It's encouraging that individual false messages are being systematically tracked, and this database of cases is online and searchable.

One of the most striking things about the East StratCom Task Force is how small it is, both in comparison to its main adversary and relative to a previous generation of counter-disinformation efforts. For comparison, the Russian Internet Research Agency -- just one arm of the Russian propaganda effort -- employs on the order of a thousand people . Many EU politicians and experts are concerned that the East StratCom Take Force effort is greatly underpowered.

2.2 Case study: Facebook

American Social media platforms' disinformation practices have been under intense scrutiny since the 2016 election . Facebook is particularly interesting because of its enormous scale, which both increases its impact and makes it a frequent target of organized disinformation campaigns.

Facebook says that most of the politically-themed disinformation or "fake news" on the platform is not politically motivated propaganda, but financially motivated clickbait , produced to drive users offsite and generate ad revenue . Regardless of motivation, Facebook has responded to disinformation in three general ways: identifying false content, identifying untrustworthy sources, and identifying "inauthentic" accounts.

Since December of 2016, Facebook users have been able to report an article as "fake news." Facebook also maintains machine learning models that try to identify posts as disinformation, including links, photos, and videos. However identified, these items are then sent to third party fact checkers for verification. These fact checkers are selected from those certified under Poynter's International Fact Checking Code of Principles and include well-known organizations such as Politifact. When a fact checker rates an article as false, it is demoted in the news feed. Regardless of rating, Facebook displays the resulting fact check as a "related article" whenever the original link is shared

Most of what is publicly known about this program comes from interviews with the fact checking organizations involved . In the U.S., Facebook currently works with six fact-checking partners who receive candidate disinformation URLs through a web "dashboard." At any one time there are perhaps a few thousand posts on the dashboard. Properly checking a claim is a meticulous and often slow research process, and one organization reported that it checked 90 stories in one year. It is unclear if the overall scale of the effort has any appreciable impact on user exposure to disinformation.

Perhaps because of this potential scale mismatch, Facebook has also started tackling disinformation at the level of sources, not just individual posts. A 2018 survey asked users "how much do you trust each of these domains?" and the resulting scores are used to "inform ranking". Again, there is no public information on the results of this program. It's also not clear how well user trust ratings align with the truth or falsity of sources, though perhaps surprisingly there may not be much partisan difference in American users' source rankings .

Meanwhile, there is an active effort at Facebook to identify and remove "inauthentic" accounts and pages, which involve a false identity or deceptive activities, and not necessarily false content. The efforts of the Russian Internet Research Agency (IRA) during the 2016 election are the best documented example of this sort of "inauthenticity." The IRA created approximately 470 Russian-controlled pages on divisive topics, with titles such as "Blacktivist", "Secured Borders," and "LGBT United". The true operators of these pages were hidden in a variety of ways, including fake accounts used to administer the page.

2.3 Case study: the Chinese information regime

21st century China provides an in-depth study of the possibilities of state information control in the networked era. Clearly the Chinese Communist Party has a particular view of what constitutes "disinformation," but consider that some kinds of Chinese disinformation are easily recognizable to Westerners. There are commercial scams of various sorts and also rumors, like the viral misconception that iodized salt could prevent radiation poisoning from Fukushima fallout, which caused a buying panic.

Chinese control of online information dissemination has tightened considerably since President Xi Jinping took power in 2013. Creemers has written a thorough review . For a few years peaking at about 2012, Chinese "microblogging" services such as Weibo provided a relatively uncontrolled public conversation. Many celebrities and other high profile people had millions, in some cases tens of millions of followers. But then a number of widely shared viral messages painted the regime in unflattering ways. Sometimes the claims were true, as when citizens collected and posted evidence of corrupt officials, and sometimes they were false, as with the claim that after the 2011 Wenzhou train crash the government "had paid more than 200 million Yuan in compensation for a foreign passenger who died." The original poster of this falsehood was sentenced to three years' imprisonment .

The Xi government decided they had a new information battleground, and made a number of moves over the next few years. Online sites and platforms must be licensed by the government. Platforms are legally responsible for their users' posts, in stark contrast to the U.S. legal framework . A Real Names law ensures that platforms and apps can link users to offline identities. If a post found to be "unlawful" is shared 500 times the poster faces up to three years in prison. A 2014 report claims two million people are involved in online content monitoring, across 800 organizations. And the "fifty cent party" of paid pro-government accounts posts an estimated 450 million comments per year.

The government has long created and disseminated its own narratives through state media organs. The so-called "50c party" (meaning "fifty cents per post") is a different approach: a network of pro-party social media accounts which are not publicly connected to the government. Yet a trove of leaked 50c worker reports shows that most posters are in fact local government employees, who often have unrelated day jobs while posting on the side . But the posts do not directly engage disinformation. King, Pan, and Roberts analyze the leaked reports to conclude, almost none of the Chinese government's 50c party posts engage in debate or argument of any kind. They do not step up to defend the government, its leaders, and their policies from criticism, no matter how vitriolic; indeed, they seem to avoid controversial issues entirely. Instead, most 50c posts are about cheerleading and positive discussions of valence issues. We also detect a high level of coordination in the timing and content in these posts. A theory consistent with these patterns is that the strategic objective of the regime is to distract and redirect public attention from discussions or events with collective action potential.

3 A Taxonomy of Tactics

Each of the institutions in these case studies has adopted specific tactics for fighting disinformation. I've collected and classified these to produce the taxonomy in Table 1. Each of these strategies will be discussed in turn, with particular attention to which are effective or desirable within a democracy. Centralized media control is not usually possible with a free press, and heavy handed strategies may conflict with democratic values such as free speech. Again, there is no attempt here to say which sorts of narratives are worthy of countering. No matter what is decreed to be disinformation, there is still the ethical question of which means of countering it best align with democratic values.

Table 1: Counter-disinformation strategies used by the three institutions in this paper, and their effectiveness and legitimacy in a democratic society.

Strategy	Used by	Effectiveness	Legitimacy
Refutation	EU Stratcom Facebook via fact-checkers	Works if consistent, but not all disinfo is about facts.	Generally legitimate to speak the truth, though people will disagree on what truth is.
Expose inauthenticity	EU Stratcom Facebook	Discredits the source, provides justification for further measures.	Content-neutrality is appealing. Important to preserve legitimate anonymity.
Alternative narratives	EU Stratcom China	Helps displace disinfo, inoculates against it if seen first.	Can itself be disinfo or distraction.
Algorithmic filter manipulation	Facebook China via 50c party	Media algorithms have huge effect on information exposure.	Platforms may abuse this power, users may game it.
Speech laws	Facebook enforces such laws China	Can be effective at targeting narrow categories of speech.	Broad laws against untruth are draconian.
Censorship	China	Effective when centralized media control is possible.	Generally conflicts with free speech.

3.1 Refutation

Refutation, rebuttal, or debunking might be the most obvious counter-strategy. It's also well within the bounds of democracy, as it's simply "more speech."

It may not be the most effective, because reasoned counter-arguments don't necessarily change someone's belief . While some experiments have shown a "backfire" effect where attempted corrections entrench false beliefs, more recent research suggests that backfire is less common than previously thought and that refutation does work to correct factual knowledge ("oh, Clinton was wrong, truck drivers do not pay more taxes than hedge fund managers…") while being unlikely to change political attitudes ("…but I'm still voting for her.") It's most effective if it's done consistently over the long term and in any case it's practiced by most counter-disinformation organizations.

The East StratCom Task Force states that it operates primarily by refuting "stories contradicting publicly available facts." This same "non-factual" claim is the moral heart of journalistic fact-checking, and Facebook's practice of demoting articles rated false. Alas, the real world is more complex than true or false. East StratCom's database of disinformation cases includes many items that are factually correct, but omit information in order to suggest a different meaning (Wardle classifies this as "misleading content"). Such messaging is more difficult to refute as simply “false.”

Interestingly, China does not engage with government critics, at least not on social media. Non-response is also a classic PR strategy. This suggests that public refutation is not a necessary tactic, and can even be harmful.

3.2 Expose inauthenticity

One of the oldest and best-recognized forms of disinformation is pretending to be someone you are not. Bot networks, "astroturfing," and undisclosed agendas or conflicts of interest could all be considered inauthentic communication. The obvious response is to discredit the source by exposing it.

China requires all apps and services to register users under their full legal names, which could be considered a pre-emptive defense against misattribution. Yet there are democratically important uses for anonymity. Insisting on traceable identities can be a severe security problem for human rights workers, journalists, and activists. Facebook says that this is why it does not require page administrators to publicly disclose their identity.

However, Facebook still knows who administers a page, and they use this information to help detect "inauthenticity." According to the company, the divisive pages created by the Russian IRA were not shut down not because of what they said -- after all, they often repeated memes and quotes from people who were genuinely involved in various political issues. Rather, they were removed because the administrators were hiding their true identities as Russian agents. Similarly, Facebook removed 30,000 French "sock puppet" or "bot" accounts which they identified "by analyzing the inauthenticity of the account and its behaviors, and not the content the accounts are publishing" and continues to remove accounts involved in “coordinated inauthentic behavior” in many countries .

The authenticity approach is morally and politically appealing because it rests on a widely shared communicative norm: pretending to be someone you are not is unethical. This avoids the difficult question of deciding what someone can be allowed to say. This idea of "content neutrality" is a key concept in U.S. First Amendment law, referenced in several Supreme Court decisions . Critics charge that a “content neutrality” doctrine is fundamentally incoherent because speech policies are never truly neutral, and can always be shaped to target particular content indirectly. Indeed, in crafting policies and directing enforcement, Facebook is exercising considerable discretion in what sort of speech to allow.

3.3 Alternative narratives

A long line of experimentation suggests that merely saying that something is false is less effective than providing an alternative narrative , and the non-platforms in this paper combat disinformation in part by promoting their own narrative.

One of the East StratCom Task Force's primary missions is to represent EU actions and values to its target audiences . This provides a ready-made positive narrative which may “inoculate” citizens against disinformation, or even displace it. The Chinese 50c party also promotes alternative narratives, albeit while pretending to be regular citizens as opposed to government workers.

At the nation-state level, alternative narratives could easy wander into suspicious territory. For this reason, much of the US government is prohibited by law from trying to influence domestic opinion. Public messaging is an essential part of a government's role, but it should not itself be disinformation.

3.4 Algorithmic filter manipulation

The rise of platforms creates a truly new way of countering disinformation: demote it by decreasing its ranking in search results and algorithmically generated feeds. Conversely, it is possible to promote alternative narratives by increasing their ranking.

On Facebook, items judged as false "appear lower in News Feed" and “typically lose 80 percent of their traffic." Users ratings of source trustworthiness similarly “inform ranking”.

While platform operators have the power to directly punish disinformation, there are other strategies for manipulating search rankings. High volume alternative narratives, as in the case of the 50c party, may be interpreted by algorithmic ranking systems as “popular” or “trending” and thereby displace other content from the resulting recommendations. This can be thought of as a technical means of distraction.

Because of their vast scale, platforms must rely on users to help moderate content by flagging spam, pornography, incitement to violence, and other undesirable posts. Facebook's disinformation detection system includes user reports . Such reporting systems can be influenced by organized campaigns, which have sometimes been able to demote content by directing an allied audience to report it, down-vote it, etc., a form of "user-generated censorship" . But ignoring audience ratings entirely is not a good answer. A distributed, bottom-up campaign to bring algorithmic attention to some issue may be a perfectly legitimate and democratic tactic. This puts platforms in the difficult position of deciding which types of collective action to reward and which to punish.

3.5 Speech laws

While there are U.S. based fake news factories targeting domestic audiences I do not know of a case where a contemporary disinformation creator was prosecuted. This may not even be possible because the Supreme Court has held that the First Amendment generally protects lying; the major exceptions concern defamation and fraud . In Europe, the recent

report of the High Level Expert Group on Fake News and Online Disinformation recommended against attempting to regulate disinformation .

China has a different approach, where people who have posted widely-shared false information on social media have been fined or imprisoned. China has also imposed licensing requirements on online platforms which publish "news," and makes platforms liable for the posts of their users. These policies create a credible threat which forces platforms to fight falsehoods (and politically unacceptable stories.)

By contrast, platforms in the U.S. operate under the legal framework of section 230 of the Communications Decency Act and are not generally responsible for the speech of their users. A number of experts and scholars feel that this is part of the reason there has been so much American innovation in social media . But in most democracies platforms are still legally liable for hosting certain types of content. For example, Germany requires platforms to remove Nazi-related material within 24 hours or face fines . These sorts of targeted regulations can be successful in the sense of largely removing certain narrow categories of posts from platforms.

3.6 Censorship

One way of combatting disinformation is simply to remove it from public view. In the 20th century, censorship was sometimes possible through control over broadcast media. This is difficult with a free press, and it is even harder to eliminate information from a networked ecosystem. Yet platforms do have the power to remove content entirely and often do, both for their own reasons and as required by law.

Censorship is generally considered suspect in a democracy. The EU's High Level Expert Group on Fake News and Online Disinformation holds that "any form of censorship either public or private should clearly be avoided."

There is no such aversion in China, which employs a variety of interlocking censorship measures. The Great Firewall prevents domestic access to offending foreign sources. Domestic platforms and publishers are licensed and held responsible for the content they distribute, which creates a self-censorship regime. Indeed, every Chinese platform has an army of people whose job it is to monitor user content for falsehoods and politically sensitive speech . WeChat has an ever-changing list of words that cannot be used in a group chat .

American platforms also employ huge numbers of moderators tasked with removing things like pornography, incitements to violence, copyright violations, and certain other material as required by law, but not typically falsehoods or political material . Outside of such specific categories, most Western platforms attempt to be "content neutral" and do not remove user material. This does not mean they do not address disinformation. Facebook does not remove content marked as false by outside fact checking organizations, but it does present that marking to its users. This seems more in line with democratic ideals.

4 Conclusion

Despite their differences, there are many common patterns between the East StratCom Task Force, Facebook, and the Chinese government. I've grouped these into six broad categories of contemporary counter-disinformation tactics, summarized in Table 1. Each has certain advantages and disadvantages in terms of efficacy and legitimacy -- that is, alignment with the values of an open society.

A cross-sector response -- both distributed and coordinated -- is perhaps the biggest challenge. In societies with a free press there is no one with the power to direct all media outlets and platforms to refute or ignore or publish particular items, and it seems unlikely that people across different sectors of society would even agree on what is disinformation and what is not. In the U.S. the State Department , the Defense Department , academics , journalists , technologists and others have all launched their own more-or-less independent counter-disinformation efforts. Distributed work is essential, but coordination is an operational advantage. In many countries, a coordinated response will require coming to terms with a deeply divided population. Citizens will require strong assurances that the strategies employed to counter disinformation are both effective and aligned with democratic values.

REFERENCES

[1] B. Adair, C. Li, J. Yang and C. Yu, "Progress Toward "the Holy Grail": The Continued Quest to Automate Fact-Checking," in *Computation + Journalism Symposium*, Evanston, 2017.

[2] E. Porter and T. Wood, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Political Behavior*, January 2018.

[3] S. C. Woolley and P. N. Howard, *Computational Propaganda Worldwide: Executive Summary*, Oxford Internet Institute, 2017.

[4] C. Paul and M. Matthews, "The Russian "Firehose of Falsehood" Propaganda Model Why It Might Work and Options to Counter It," RAND Perspectives, 2016. [Online]. Available: https://www.rand.org/pubs/perspectives/PE198.html.

[5] H. Romerstein, "Counterpropaganda: We can't win without it," in *Strategic Influence: Public Diplomacy, Counterpropaganda, and Political Warfare*, J. M. Waller, Ed., Institute of World Politics Press, 2009, pp. 137-180.

[6] N. J. Cull, The Cold War and the United States Information Agency: American Propaganda and Public Diplomacy, 1945-1989, Cambridge University Press, 2009.

[7] M. Hellman and C. Wagnsson, "How can European states respond to Russian information warfare? An analytical framework," *European Security*, vol. 26, no. 2, 2017.

[8] "Questions and Answers about the East StratCom Task Force," European External Action Service, 12 May 2018. [Online]. Available: https://eeas.europa.eu/headquarters/headquarters-homepage_en/2116/%20Questions%20and%20Answers%20about%20the%20East%20StratCom%20Task%20Force.

[9] J. Rankin, "https://www.theguardian.com/world/2017/nov/25/eu-anti-propaganda-unit-gets-1m-a-year-to-counter-russian-fake-news," *The Guardian*, 17 November 2017.

[10] C. McDonald-Gibson, "The E.U. Agency Fighting Russia's Wildfire of Fake News with a Hosepipe," *Time*, 11 September 2017.

[11] EU Vs. Disinfo, "Figure of the Year: 1310," 20 December 2017. [Online]. Available: https://euvsdisinfo.eu/figure-of-the-year-1310/.

[12] EU Vs. Disinfo, "Disinformation Cases," [Online]. Available: https://euvsdisinfo.eu/disinformation-cases/. [Accessed 13 1 2019].

[13] T. Lister, J. Sciutto and M. Ilyushina, "Putin's 'chef,' the man behind the troll factory," 17 October 2017. [Online]. Available: https://www.cnn.com/2017/10/17/politics/russian-oligarch-putin-chef-troll-factory/index.html.

[14] N. MacFarquhar, "Inside the Russian Troll Factory: Zombies and a Breakneck Pace," 18 February 2018. [Online]. Available: https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html.

[15] M. Maksimovic, "EU officials warn of 'underestimating' Russian propaganda in Balkans," *Deutsche Welle*, 14 November 2017.

[16] M. Isaac and D. Wakabayashi, "Russian Influence Reached 126 Million Through Facebook Alone," *New York Times*, 30 October 2017.

[17] E. Schrage, "Hard Questions: Russian Ads Delivered to Congress," 2 October 2017. [Online]. Available: https://newsroom.fb.com/news/2017/10/hard-questions-russian-ads-delivered-to-congress/.

[18] L. Moses, "'The underbelly of the internet': How content ad networks fund fake news," 28 November 2016. [Online]. Available: https://digiday.com/media/underbelly-internet-fake-news-gets-funded/.

[19] A. Mosseri, "News Feed FYI: Addressing Hoaxes and Fake News," Facebook, 15 December 2016. [Online]. Available: https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

[20] A. Woodford, "The Hunt for False News," Facebook, 19 October 2018. [Online]. Available: https://newsroom.fb.com/news/2018/10/inside-feed-hunt-false-news-october-2018/.

[21] Poynter Institute, "The International Fact-Checking Network," 2017 September. [Online]. Available: https://www.poynter.org/channels/fact-checking.

[22] D. Zigmond, "Machine Learning, Fact-Checkers and the Fight Against False News," Facebook, 8 April 2018. [Online]. Available: https://newsroom.fb.com/news/2018/04/inside-feed-misinformation-zigmond/.

[23] J. Smith, "Designing Against Misinformation," 20 December 2017. [Online]. Available: https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.

[24] M. Ananny, "The partnership press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation," Tow Center for Digital Journalism, 4 April 2018. [Online]. Available: https://www.cjr.org/tow_center_reports/partnership-press-facebook-news-outlets-team-fight-misinformation.php.

[25] A. Mosseri, "Helping Ensure News on Facebook Is From Trusted Sources," Facebook, 19 January 2018. [Online]. Available: https://newsroom.fb.com/news/2018/01/trusted-sources/.

[26] G. Pennycook and D. G. Rand, "Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality," *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, 2019.

[27] J. Weedon, W. Nuland and A. Stamos, *Information Operations and Facebook*, Facebook, 2017.

[28] C. Timberg, "Russian propaganda may have been shared hundreds of millions of times, new research says," The Washington Post, 5 October 2017. [Online]. Available: https://www.washingtonpost.com/news/the-switch/wp/2017/10/05/russian-propaganda-may-have-been-shared-hundreds-of-millions-of-times-new-research-says.

[29] R. Creemers, "Cyber China: Upgrading Propaganda, Public Opinion Work and Social Management for the Twenty-First Century," *Journal of Contemporary China*, vol. 26, no. 103, 2017.

[30] L. Burkitt, "Fearing Radiation, Chinese Rush to Buy...Table Salt?," *The Wall Street Journal*, 17 March 2011.

[31] C. Yim, "Rumors Confession," 12 April 2014. [Online]. Available: https://www.pressreader.com/china/china-daily/20140412/281530814003426.

[32] The Electronic Frontier Foundation, "CDA 230: The most important law affecting Internet speech," [Online]. Available: https://www.eff.org/issues/cda230. [Accessed 31 12 2017].

[33] M. Fong, "China Monitors the Internet and the Public Pays the Bill," *Global Voices Advox*, 20 July 2014.

[34] G. King, J. P. Pan and M. E. Roberts, "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument," *American Political Science Review*, vol. 111, no. 3, 2017.

[35] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz and J. Cook, "Misinformation and Its Correction Continued Influence and Successful Debiasing," *Psychological Science in the Public Interest*, vol. 13, no. 3, 2012.

[36] M. Chessen, "Understanding the Psychology Behind Computational Propaganda," Advisory Commission on Public Diplomacy, May 2017. [Online]. Available: https://uscpublicdiplomacy.org/story/can-public-diplomacy-survive-internet.

[37] C. Silverman, "The Backfire Effect," *Columbia Journalism Review*, 17 June 2011.

[38] E. Porter, "Facts matter, and people care: An empirical perspective," Advisory Commission on Public Diplomacy, May 2017. [Online]. Available: https://uscpublicdiplomacy.org/story/can-public-diplomacy-survive-internet.

[39] EU vs. Disinfo, "COMMENTARY: Means, goals and consequences of the pro-Kremlin disinformation campaign," 19 January 2017. [Online]. Available: https://euvsdisinfo.eu/commentary-means-goals-and-consequences-of-the-pro-kremlin-disinformation-campaign.

[40] C. Wardle, "Fake news. It's complicated.," *First Draft News*, 16 February 2017.

[41] N. Gleicher, "Coordinated Inauthentic Behavior Explained," Facebook, 6 December 2018. [Online]. Available: https://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.

[42] E. Chemerinsky, "Content Neutrality as a Central Problem of Freedom of Speech: Problems in the Supreme Court's Application," *Southern California Law Review*, vol. 74, 2001.

[43] J. Henick and R. Walsh, "U.S. 2016 Elections: A case study in "inoculating" public opinion against disinformation," Advisory Commission on Public Diplomacy, May 2017. [Online]. Available: https://uscpublicdiplomacy.org/story/can-public-diplomacy-survive-internet.

[44] European External Action Service, "Questions and Answers about the East StratCom Task Force," 8 November 2017. [Online]. Available: https://eeas.europa.eu/headquarters/headquarters-homepage_en/2116/%20Questions%20and%20Answers%20about%20the%20East%20StratCom%20Task%20Force.

[45] W. R. Sager, "Apple Pie Propaganda: The Smith-Mundt Act before and after the Repeal of the Domestic Dissemination Ban," *Northwestern University Law Review*, vol. 109, no. 2, 2014.

[46] T. Lyons, "News Feed FYI: Replacing Disputed Flags with Related Articles," Facebook, 20 December 2017. [Online]. Available: https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

[47] C. Peterson, "User-generated censorship : manipulating the maps of social media," June 2013. [Online]. Available: http://hdl.handle.net/1721.1/81132.

[48] L. Sydell, "We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned," *NPR*, 2016 November 23.

[49] T. McCoy, "Inside a Long Beach Web operation that makes up stories about Trump and Clinton: What they do for clicks and cash," 16 November 2016. [Online]. Available: http://www.latimes.com/business/technology/la-fi-tn-fake-news-20161122-story.html.

[50] E. Cherminsky, "The First Amendment and the Right to Lie," 12 September 2012. [Online]. Available: http://www.abajournal.com/news/article/the_first_amendment_and_the_right_to_lie.

[51] "Final report of the High Level Expert Group on Fake News and Online Disinformation," 12 March 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.

[52] M. Eddy and M. Scott, "Delete Hate Speech or Pay Up, Germany Tells Social Media Companies," 30 June 2017. [Online]. Available: https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html.

[53] "'It's seen as a cool place to work' – how China's censorship machine is becoming a growth industry," 29 September 2017. [Online]. Available: http://www.scmp.com/news/china/policies-politics/article/2113377/its-seen-cool-place-work-how-chinas-censorship-machine.

[54] L. Ruan, J. Knockel, J. Q. Ng and M. Crete-Nishihata, "One App, Two Systems: How WeChat uses one censorship policy in China and another internationally," CitizenLab, 30 November 2016. [Online]. Available: https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/.

[55] A. Chen, "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed," *Wired*, 23 October 2014.

[56] U.S. Department of State, "Global Engagement Center," [Online]. Available: https://www.state.gov/r/gec/. [Accessed 1 1 2018].

[57] U.S. Department of Defense, "Strategy for Operations in the Information Environment," June 2016. [Online]. Available: https://www.defense.gov/Portals/1/Documents/pubs/DoD-Strategy-for-Operations-in-the-IE-Signed-20160613.pdf.

[58] A. Marwick and R. Lewis, "Media Manipulation and Disinformation Online," [Online]. Available: https://datasociety.net/output/media-manipulation-and-disinfo-online/.

[59] L. H. Owen, "The Trust Project brings news orgs and tech giants together to tag and surface high-quality news," 16 November 2017. [Online]. Available: http://www.niemanlab.org/2017/11/the-trust-project-brings-news-orgs-and-tech-giants-together-to-tag-and-surface-high-quality-news/.

Misinfosec

Applying Information Security Paradigms to Misinformation Campaigns

Christopher R. Walker
Marvelous AI
San Francisco, CA, USA,
walker@marvelous.ai

Sara-Jayne Terp
SOFWERX
Tampa, FL, USA,
sarajterp@gmail.com

Pablo C. Breuer
U.S Special Operations Command
Tampa, FL, USA,
Pablo.Breuer.socom.mil

Courtney L. Crooks PhD
Georgia Tech Research Institute (GTRI)
Atlanta, GA, USA, courtney.crooks@gtri.gatech.edu

ABSTRACT

State actors, private influence operators and grassroots groups are all exploiting the openness and reach of the Internet to manipulate populations at a distance, extending their decades-long struggle for "hearts and minds" via propaganda, influence operations and information warfare. Computational propaganda fueled by AI makes matters worse.

The structure and propagation patterns of these attacks have many similarities to those seen in information security and computer hacking. The Credibility Coalition's MisinfosecWG working group is analyzing those similarities, including information security frameworks that could give the truth-based community better ways to describe, identify and counter misinformation-based attacks. Specifically, we place misinformation components into a framework commonly used to describe information security incidents. We anticipate that our work will give responders the ability to transfer other information security principles to the misinformation sphere, and to plan defenses and countermoves .

CCS CONCEPTS

• Security and privacy~Trust frameworks • Security and privacy~Social aspects of security and privacy • General and reference~Reliability

KEYWORDS

infosec, misinformation, disinformation, viral deception, influence campaigns

ACM Reference format:

Christopher R. Walker, Sara-Jayne Terp, Pablo C. Breuer, Courtney L, Crooks, PhD. 2019. Misinfosec: Applying Information Security Paradigms to Misinformation Campaigns. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316742

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316742

1 Introduction

1.1 Misinformation

There is no coherent and widely-adopted definition of *misinformation* (or of *truth*, *credibility*, *trust*, etc).

Disinformation is sometimes defined as dissemination of explicitly false or misleading information; and misinformation the communication of false information without intent to deceive, manipulate or otherwise obtain an outcome. Elsewhere, *misinformation* is used for the deliberate case, both inclusive and exclusive of the accidental case. While we think that definitions are important, we suspect that the common uses of misinformation (and disinformation) hamper their use as technical terms drawing a distinction between intentional and accidental spread. While attacks are inherently intentional, the spread of false information by individuals within an attack may not be. We will not explicitly use terminology to draw this distinction here.

We use *misinformation attack* (and *misinformation campaign*) to refer to *the deliberate promotion of false, misleading or mis-attributed information.* Whilst these attacks occur in many venues (print, radio, etc), we focus on the creation, propagation and consumption of misinformation online. We are especially interested in misinformation designed to change beliefs in a large number of people.

Actors behind misinfo attacks include nation-states, institutional actors, grassroots trolls and financially-motivated freelancers. Common motives include geopolitical aims, issue-promotion, or financial gain. In the run-up to the 2016 US election, websites churning out fabricated stories were a cottage industry. Governments worldwide are studying misinformation as a form of influence operation or information war.

1.2 Misinformation work lacks frameworks

Analysts and engineers have been creatively adapting their own techniques to a range of issues that they perceive to be essential to confronting the larger problem of online misinformation. Which sites promote false stories? How can you detect a false story? How can you detect a false statement? How can you detect a doctored photograph? How can you detect a sock-puppet or a troll? Some platforms have processes in place for detecting and mitigating nefarious user activity. And independent technical approaches have also begun to sprout up. These techniques and toolkits will have potential application in the broader domain of combating misinformation. But tools in themselves have no values and serve no inherent masters.

Misinformation operates within a complicated socio-technical ecosystem, so the approach must be multidisciplinary. Just as cyber security professionals must keep up with constantly evolving techniques and strategies for exploiting private information available within cyberspace, misinformation professionals should expect the same. Further, policy makers must be enabled to develop and enforce data privacy policies that help protect their public from nefarious online activity and privacy breaches. The application of technical and operational techniques can only be productive in the context of thoughtfully designed tactics, grounded in well-defined strategy. Worse, absent strategic and tactical goal-setting we cannot be certain that these techniques are either necessary or sufficient to the amelioration of the threat presented by online influence operations. Without a framework, we are just stabbing in the dark.

1.3 MisinfoSec working group (MisinfosecWG)

The *MisinfoSec Working Group (MisinfosecWG)* is part of the Credibility Coalition, which is an interdisciplinary community committed to addressing the proliferation and amplification of misinformation online, through transparent and collaborative research.

Combining "misinformation" and "information security," MisinfosecWG is developing a framework for understanding organized misinformation attacks based on existing information security principles. Specifically, we promote a more formal and rigorous treatment of 1) detecting misinformation-based attacks and 2) protecting against misinformation-based attacks.

1.4 Existing work on misinfosec

This work grew out of earlier work describing red team and blue team misinformation tactics and characterizing misinformation as an information security problem that infosec frameworks and principles could be applied to .

Misinformation is slowly becoming a subject of interest to information security teams. FireEye helped characterize the Iranian IUVM disinformation network, tracked disinformation typosquats and analyzed traffic in the 2018 US midterm elections; ThreatConnect tracked online infrastructure behind Russian misinformation campaigns, and Synack described how misinformation could be used as part of an information security attack.

The extension of information security to include misinformation is also under discussion. Landau argues that the NSPD-54 definition of cybersecurity should be extended to include information operations (e.g. misinformation), and raises the issue of misinformation users adapting their tools and techniques as detection improves. Rogers frames misinformation as an information integrity problem, citing the infosec concept of maintaining CIA (*confidentiality*, *integrity*, *availability*), and suggests applying infosec practices such as threat modelling. Brockman and Grugq describe parallels between misinformation and information security. Lin and Kerr examine cyber-enabled information warfare as a conflict form where the USA is weak, examine the environment, operations and characteristics of this space, and call for new tactics and responses in it.

1.5 Methodology

Our methodology is to analyze known misinformation attacks to identify their components. What are the atomic *actions* in propaganda attacks? How do actions combine to form larger events, including more complex actions and *attacks*? How do the instances of attacks and actions combine to form *campaigns*?

We then place those components into a framework (e.g. ATT&CK) commonly used to describe information security incidents. The outputs from the group include a misinfosec threat matrix designed for use by "Blue Teams" considering options for defense and counter-attack, and by "Red Teams" anticipating future attack types.

2 MisinfosecWG Activities

2.1 Analyzing Misinformation Campaigns

MisinfosecWG is analyzing known misinformation campaigns. A *campaign* is online manipulation designed to influence the beliefs of a large number of people. It typically consists of several attacks or incidents, aligned toward a specific goal.

In 2017, at least 18 countries used misinformation tactics in elections. Most employed groups of "opinion shapers" to manipulate domestic elections; some, including Russia and Iran, used these tactics to manipulate popular beliefs in other countries. The more well-known campaigns include Russian interference and influence on Brexit, the 2017 French Presidential election, the election of Donald Trump, attacks on the Parkland teenagers, promotion of Jade Helm conspiracy theories and various influence operations around the Black Lives Matter movement. And there are countless less well-known international operations. Private influence operators manipulate beliefs; grassroots 'trolls' and marketers use misinformation campaigns to push agendas or make money, usually from online advertising.

With few exceptions, most of the response thus far has been akin to whack-a-mole. The extent of the threat and the range of

possible actions are simply not understood well enough to formulate counter-moves, whether tactical or strategic.

In our analysis, we look at the actors and their presumed goals and timeframes. We also: look in detail at the methods used in each attack; look at the counters used against them; and list related attacks, as a first pass at creating attack types that can be grouped and discussed together.

2.2 Components of Influence Campaigns

Benkler et al provide a number of useful distinctions. First, they suggest viewing categories of online information threats in terms of a few scalar dimensions, such as *centralized* versus *decentralized*, *political* versus *commercial* and *technological* versus *institutional*. These dimensions suggest we focus on factors such as the *actors*, *objectives*, and *delivery mechanisms* in describing the terrain.

*2.2.1 Actors and Objectives.:*Influence operations are undertaken by an *attacker*, directed at a *target* and sometimes amplified by *carriers*. These actors can all be individuals, populations or institutions. But ultimately the targets are people, or aggregates of people. The objective is typically tied up in the psychology of the actors, especially the targets and carriers. The goal is to change the beliefs and behavior of individual people, often at scale, via manipulation: *directly influencing beliefs, attitudes, or preferences of a target population in ways that are not normatively appropriate in context.* The cognitive objectives of the operations often include widespread factual misunderstanding or confusion. This can vary from *gaslighting* and *disorientation* to *distraction, priming* and *agenda setting*.

*2.2.2 Message, Delivery and Propagation.:*The tactical objective of an attack is typically cognitive or social. The payload is information. Misinformation is semantically misleading, contextually misleading, misattributed, or factually incorrect. The information is transferred via communication events, which themselves can be varied and subtle; some are harder to address than others. The information can range from propaganda to bullshit. *The bullshit artist "does not care whether the things he says describe reality correctly. He just picks them out, or makes them up, to suit his purpose."*

2.3 Social Network Architecture and Message Contagion

The social factors at play involve pathways of communication and attention. For the most part, this amounts to social media. But the interaction between "mainstream" propagandists and their social media sock-puppets and supporters is highly textured. Benkler refer to this interaction as *network propaganda*, noting the familiarity effect with this approach.

Moreover, the types of informational networks that serve as a substrate for the propagation of these messages vary significantly. Some networks are *truth-corrective*: they sanction information that diverges from verified reporting. Other networks are *narrative-corrective*: they sanction information that diverges from the consensus narrative of the network.

And the role of these network-driven attention frameworks on the promotion of marginal framings into the mainstream should not be underestimated. Benkler also describes an attention backbone which promotes stories from the periphery of the network and *propaganda feedback loops* which are *pathological network dynamics in which (mostly attentional) sanctions are imposed for breaking with the received narrative preferred by the target population.*

In short, the objectives are cognitive and the vectors of delivery are social. Successful responses to attacks must be crafted in the context of these varied actors, targets, messages, goals and networks.

2.4 Why Information Security?

We considered adapting frameworks from several fields. This included advertising frameworks, lean enterprise frameworks and information security (infosec) frameworks. We chose the infosec framework because of the close fit between infosec attacks on individual and networks of machines and misinformation attacks on individuals and networks of humans.

The most frequently mentioned alternative is advertising. Tactics and techniques will inevitably draw from numerous fields, advertising prominent among them. For example, advertising test cases might be able to tell us something about the psychosocial motivators that are most effective in advertising. In other words, what techniques are effective in getting people interested or invested in the "product" of the advertisement? Understanding these motivators could be helpful in understanding what captures people's attention in misinformation campaigns.

But while advertising may have much to offer at the tactical level, we ultimately decided that it did not offer an adequate fit to our problems as a framework. Most notably, advertising uses a subset of the attack techniques that we care about, but does not typically convince an audience to do something against their own interest nor does it fabricate false facts. Furthermore, a contrarian might note that tactics from the advertising domain would more appropriately be drawn from the world of anti-corporate advocacy. Efforts to ameliorate ad placement and saturation would have quite a bit to say about which responses are most effective, whether regulatory or otherwise.

Similarly, scholars of health misinformation could be a significant source of tactical insight, given the recent Measles outbreak and its relation to misinformation regarding vaccinations. Our approach is inter-disciplinary, but our framework is information security.

2.5 Adapting Information Security Frameworks

Information security encompasses offensive and defensive computer network operations, electronic warfare, psychological operations, military deception, and operational security. Information security is a robust field with well-understood principles and best practices, covering physical, informational and cognitive dimensions of the information environment.

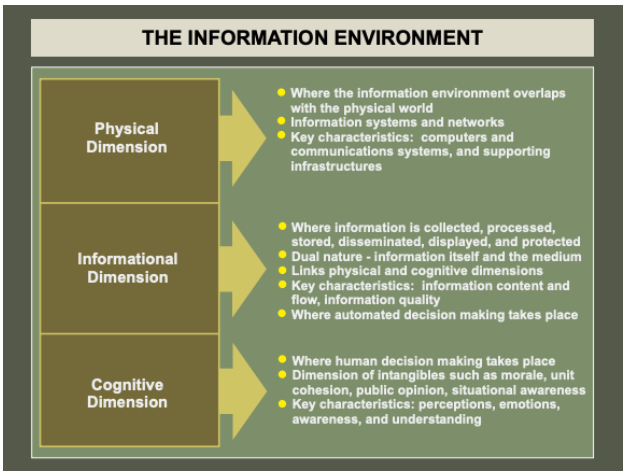


Figure 1: Dimensions of The Information Environment

Alerting systems exist on top of frameworks and standards describing information attacks like DDOS, viruses, and unwanted internet traffic like spam etc. These systems could be a good place to start with misinformation.

We explored potential fits between misinformation and several common information security frameworks: strategic-level models like the *SANS sliding scale* (architecture, passive defense, active defense, intelligence, offense), *Gartner cycle* (prevent - detect - respond - predict), *NIST framework* (detect - protect - identify - recover - respond) and *Cyber Attack Lifecycle*, and operational-level models like the *ATT&CK matrix* and *SANS top 20* .

The Cyber Attack Lifecycle basically maps to our campaign descriptions (*reconnaissance*, *weaponization*, *installation*, *exploitation*, *command-and-control*, and *actions on the objective*). We could also use this lifecycle to link goals and intent (e.g. the "four Ds" of propaganda: Dismiss, Distort, Distract, Dismay).

The MITRE ATT&CK Matrix covers the last three stages of the cyber attack lifecycle, and lists tactic phases (initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, exfiltration, command and control) with a set of techniques that an adversary could use in each phase (e.g. *Spearfishing Attachment* is a type of *Initial Access* tactic). The ATT&CK database provides examples, detection and mitigation for each technique, along with extensive references.

The ATT&CK Tactics, Techniques and Procedures (TTPs) describe *patterns of activities or methods associated with a specific threat actor or group of threat actors.* Tactics are the top-level steps that an attacker typically takes (e.g. "amplify message"); techniques are the different ways those steps can be done (e.g. "repeat message using bots"); procedures are the sequences of actions in an attack. ATT&CK TTPs were created by taxonomizing existing information security threat reports and analyses. A similar process of assembling and grouping TTPs could also work well in support of Blue Team efforts in Misinformation Security (MisinfoSec).

3 A strawman misinformation framework

3.1 Campaigns: Advanced Persistent Threats

In information security, an Advanced Persistent Threat (APT) is an attack or an attacker operating over a long period of time. APTs are usually (but not always) backed by nation-states. In misinformation, APTs usually run long-duration campaigns. The canonical nation-state misinformation campaign is the 2015-2017 Russian troll farm work on the 2016 US presidential elections. Jamieson describes these operations in detail. The objectives are numerous: *an amplifying effect; an agenda-setting effect; a normative effect; target identification; a mobilizing effect; a two-step flow effect; weighting, contagion and spiral of silence effects;* and *a familiarity effect.*

Benkler looked at the online spread of prominent political stories before and after the 2016 US presidential elections. While the authors discuss aspects of the Russian campaign, their primary focus is the online media ecosystem itself. First, the online political environment in the United States is polarized, but the "filter bubbles" are best characterized as 1) the Fox News bubble and 2) everyone else. Second, the corrective sanctions at play in the two environments are highly asymmetric. The latter bubble penalizes for straying from the truth while the former penalizes for straying from the accepted narrative. Interventions against misinfo attacks in these two environments could be very different.

3.2 Incidents: building-blocks of campaigns

Campaigns are often built from smaller building blocks. We will refer to those as *incidents.* One example is the 2014 Columbian Chemicals incident, which we've listed as:

Summary: Early Russian (IRA) "fake news" stories. Completely fabricated; very short lifespan.

Actor: probably IRA (source: recordedfuture)

Timeframe: Sept 11 2014 (1 day)

Presumed goals: test deployment

Artifacts: text messages, images, video

Method: 1. Create messages. e.g. "A powerful explosion heard from miles away happened at a chemical plant in Centerville, Louisiana #ColumbianChemicals" 2. Post messages from fake twitter accounts; include handles of local and global influencers (journalists, media, politicians, e.g. @senjeffmerkley) 3. Amplify, by repeating messages on twitter via fake twitter accounts

Result: limited traction

Counters: None seen. Fake stories were debunked very quickly.

Related attacks: These were all well-produced fake news stories, promoted on Twitter to influencers through a single dominant hashtag -- *#BPoilspilltsunami, #shockingmurderinatlanta, #PhosphorusDisaster, #EbolaInAtlanta*

3.3 Tactics and Techniques

A complete list of misinformation tactics will likely map well to the existing ATT&CK framework tactics. For example, a strawman for this could be as follows.

Table 1: Partial Tactics and Techniques Matrix

Example Tactic	Example Techniques
Initial access	Account takeover Create fake group Parody account Deep cover
Create artefacts	Steal existing artefacts Deepfake
Insert theme	Create fake emergency
Amplify message	Repeat messaging with bots Create fake argument Buy friends
Command and control	Create fake real-life events

This top-down analysis is also being augmented with bottom-up analysis of attacks, looking at the individual components of each attack, and examining and listing those as techniques to fit into the framework.

Table 2: Examples of Techniques and their Descriptions

id	Short	Desc
1	two-events-one-place	Organize two opposing physical events for the same time and place
2	widen-existing-rifts	Increase emotions and use non-false information to widen existing rifts
3	create-fake-emergency	Raise alarm about nonexistent emergency

This process is iterative and collaborative. We will refine these strawmen into more detailed TTP descriptions and recommendations as options are tested or eliminated.

Furthermore, we will define major terms of art at focal points on the scale, with an emphasis on descriptive or procedural rigor. One of the operating assumptions of MisinfosecWG is that social and cognitive factors can "scale up and down" -- facilitating some definitional and procedural crossover in both the construction of a framework for understanding these attacks and in their detection.

3.4 Procedures: New forms of attack

Once we have components, we can put them together in new ways and discover threat types that we might not have considered before. A checklist of threats and best practices creates the space necessary to think more strategically about the misinformation environment and to balance institutional needs in the context of well-tested security principles.

4 Potential Uses

This framework will give misinformation responders the ability to transfer other information security principles to the misinformation sphere, identify gaps in known attack types, plan defenses and countermoves to common components, assess tools and mechanisms, build an information security style alert structure (cf US-CERT) and plan defenses for the types of large-scale adaptive threats that machine learning and other automation makes possible.

Most offensive computer network operations are based upon misinformation, which should aid in our task: network intruders want their targets to make decisions or take actions advantageous to them based on information that's shown, hidden, altered or destroyed, e.g. STUXNET allegedly hid information about the true state of centrifuges from operators, enabling them to make an incorrect decision that no action needed to be taken.

In infosec, an organized taxonomy of attack and defense techniques allows operators to apply well-tested responses to familiar attack patterns, and learn from both successful and failed attacks. For misinformation, these interventions are likely to be drawn from various disciplines including sociology and psychology, e.g. one possible intervention to a misinformation campaign is to push inoculating information--new information that draws opinion away from the goal of the original misinformation. At the campaign scale, some of these interventions will need to be prepared in advance for deployment in specific, measurable contexts, e.g. the Macron teams' preparation in the 2017 French elections for the reuse of a 2016 US Presidential election technique (releasing and amplifying information from leaked political emails).

4.1 Red-Team, Blue-Team Exercises

To build a good defense, you need to understand your threat surface and the types of attacks that are likely on it. The best way to understand attacks is to attack; in information security this is done through simulated attacks, where a "red team" attacks the systems of a defending "blue team". These exercises typically expose previously-unseen system vulnerabilities.

Brundage et al outlines a first red team playbook for misinformation, with common examples of: actors, targets, payloads, objectives, automation, and techniques. Extending this with the misinfosec work will give us a detailed catalogue of attack types with corresponding blue team, allowing platform defenders, and autonomous blue team actors to test and stage effective counter-measures.

4.2 Alerting and Defense

Absent a comprehensive counter-influence strategy from major government and intelligence players, the responsibility for misinformation management currently falls on individual persons and institutions. This mirrors the history of the information security field, which created bodies like the US-CERT organisation: the US government body that coordinates defenses and responses to cyber attacks.

A similar body for sharing and alerting may be needed here. US-CERT's work includes threat monitoring and analysis, information sharing, analytics, operations, communications and international partnerships. Its outputs include a current activity list, monthly active summaries, alerts, notes and tips and security publications.

These activities and outputs map well but not exactly to misinformation (different responses and connections are needed). And US-CERT already has a sister organization, ICS-CERT, which covers security of industrial control systems. We don't yet know which organizational roles will be the end user for a new body's product. Whoever is responsible for the adoption, deployment and enforcement of these practices, will probably require help from security professionals.

4.3 Challenges involved in using frameworks from information security to misinformation

While information security attacks are firmly rooted in the quantitative field of computer science, influence campaigns are, by necessity, rooted in the qualitative fields of sociology and psychology. The linking of quantitative and qualitative fields of science has always been epistemically precarious. Additionally, any attempt to develop an overarching and generalized framework will necessarily omit details. No overarching framework will ever be completely accurate in all situations. This framework attempts to provide an ontology for influence campaigns whether or not they are executed exclusively in the cyber domain. As with any complex system, there will be an emergence of properties which is greater than the sum of its parts. Finally, this framework attempts to examine influence from the view point of tactics, techniques, and procedures (TTPs) without consideration for the intent, morality, or legality of such actions. Analysis of morality, legality, and intent are beyond the scope of this work.

4.4 Counterattack (and its limitations)

Democracies face structural disadvantages relative to the producers of misinformation. Clint Watts cites the Kremlin's strategic edge from Russia's cybercrime underworld and the plausible deniability it gives; the US and its allies don't have this advantage.

Ultimately, America's problem in counterinfluence is that we don't know what to say … During the Cold War, the United States promoted democracy and democratic values. But today the United States doesn't appear to know what it wants. Quite simply, if America doesn't have its feet on the ground, then it can't push back at those challenging us.

This seems right, but generational. In the meantime, we need a plan. One that doesn't sacrifice what we believe in. This is hard. We believe that transparency is key for democratic actors. Knowing who is ***actually*** delivering the message goes a long way. Of course, there are cases in which we want to protect the source (e.g. dissidents in an autocracy), but those are the exception rather than the rule. The framework we propose is agnostic to locality and describes components; it is up to the "local populace" to decide acceptable countermeasures. In other words, part of the remedy is democratic participation itself.

5 Conclusions and Further Work

We've started the work of adapting information security frameworks for misinformation tracking and counters, but there is much work still to do. The information security field has decades of experience that we can draw on in our work, but there have been enough differences between the fields for us to create a new framework, albeit one based on ATT&CK. Challenges we anticipate in this work include epistemology, working across multiple very different fields of research, defining and naming different levels and stages of 'attack'; persuading people that information security frameworks are already about human influence systems, and legal and ethical constraints on response.

This paper focussed on the idea of mapping misinformation to infosec frameworks. Next we complete these frameworks, to understand how we would test and implement specific defenses, before implementing and testing them. Please join us! You don't have to be an infosec person to help: we have lots of small tasks (like finding and describing classic large-scale misinformation 'attacks'). Along the way, we're hoping that more infosec people will understand misinformation better, and more misinformation people will learn about techniques we can apply from a parallel field. Join us here: https://github.com/credcoalition/community-site/wiki/How-to-Help

REFERENCES

[1] Y. Benkler, R. Farris and H. Roberts, Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics, Oxford: Oxford University Press, 2018.

[2] Smith and Banic, "Fake News: How a partying Macedonian teen earns thousands publishing lies," NBC News, 2016.

[3] R. DiResta, "THE INFORMATION WAR IS ON. ARE WE READY FOR IT?," Wired, 2018.

[4] A. Petrov, "Modeling Position Selection by Individuals during Information Warfare in Society," *Mathematical Models and Computer Simulations*, pp. 401-8, 2016.

[5] S. Oates, "When Media Worlds Collide: Using Media Model Theory to Understand How Russia Spreads Disinformation in the United States," in *American Political Science Association 2018 Annual Meeting*, Boston, MA, 2018.

[6] S. Oates, J. B. Barrow and B. Foster, "From Network to Narrative: Understanding the Nature and Trajectory of Russian Disinformation in the U.S. News.," in *International Journal of Press/Politics Conference*, Oxford, UK, 2018.

[7] A. Field, D. Kliger, S. Wintner, J. Pan, D. Jurafsky and Y. Tsvetkov, "Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies.," EMNLP, 2018.

[8] T. Boucher, "Adversarial social media tactics," Medium, 2018.

[9] S.-J. Terp, "Security frameworks for misinformation," 2018.

[10] S.-J. Terp, "Practical influence operations," 2018.

[11] S.-J. Terp, "Social engineering at scale," 2018.
[12] Stubbs, "Exclusive Iran based political influence operation - bigger, persistent, global," Reuters, 2018.
[13] F. Mortola, "Disinformation through fabricated news site," 2018.
[14] Foster, "Influence Operations Targeting the 2018 U.S. Midterms: What are we seeing? What are we not?," 2018.
[15] Ehmke, "Influencer Vaccine: Identifying Information Operations Infrastructure," 2018.
[16] Kuhr, "Leveraging threat intel disinformation campaigns to defeat attribution," 27 February 2017. [Online]. Available: https://www.synack.com/2017/02/27/shmoocon-2017-recap-election-hackers-vs-threat-intel-attribution/.
[17] Landau, "Cybersecurity: time for a new definition," 2018.
[18] Rogers, "Fake news as an information security problem," May.
[19] Rogers, Director, *Fake news as an information security problem.* [Film]. 2018.
[20] D. Gordon, "DEFENDING THE INDEFENSIBLE: A NEW STRATEGY FOR STOPPING INFORMATION OPERATIONS," 2018.
[21] Grugq., "Lessons in Cyber: Influence Operations," 2018.
[22] A. Zegart, H. Lin, T. Fingar, N. Persily and L. Ross, "Cyber-Enabled Information and Influence Warfare and Manipulation: Understanding Problems, Developing Solutions," 2017.
[23] H. Lin and J. Kerr, "On Cyber-Enabled Information/Influence Warfare and Manipulation," 2017.
[24] H. Lin, "Developing Responses to Cyber-Enabled Information Warfare and Influence Operations," 2018.
[25] FreedomHouse., "Manipulating Social Media to Undermine Democracy," 2018.
[26] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini and J. Blackburn, "Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls," 2018.
[27] D. D. Kirkpatrick, "Signs of Russian Meddling in Brexit Referendum," 2017.
[28] D. O'Sullivan and D. Byers, "Exclusive: Fake black activist accounts linked to Russian government," 2017.
[29] V. Joler, M. Jovanović and A. Petrovski, "Mapping and quantifying political information warfare Part 1 : Propaganda, domination & attacks on online media," 2016.
[30] M. Rosenberg, N. Confessore and C. Cadwalladr, "How Trump Consultants Exploited the Facebook Data of Millions," 2018.
[31] A. Nossiter, D. E. Sanger and N. Perlroth, "Hackers Came, but the French Were Prepared," 2017.
[32] H. G. Frankfurt, On Bullshit, Princeton, NJ: Princeton Univerity Press, 2005.
[33] DoD, "Joint Publication 3-13: Information Operations, February 13, 2006," 2006.
[34] Cyberpedia, "HOW TO BREAK THE CYBER ATTACK LIFECYCLE," 2019.
[35] B. Ninmo, "The 4 Ds of Propaganda: Dismiss, Distort, Distract, Dismay," 2017.
[36] MITRE, "Att&ck matrix for enterprise," 2019. [Online]. Available: https://attack.mitre.org/.
[37] "Post-Exploit Threat Modeling with ATT&CK," 2016.
[38] K. H. Jamieson, Cyberwar: How Russian Hackers and Trolls Helped Elect a President - What We Don't, Can't, and Do Know, Oxford. UK: Oxford University Press, 2018.
[39] M. Brundage, A. Shahar, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. C. Allen, J. Steinhardt and C. Flynn, "The Malicious Use of Artificial Intelligence:," Archiv, 2018.
[40] M. Brockman, "Data-Driven Propaganda as a Subset of Adversarial Examples," 2018.
[41] C. Watts, Messing with the Enemy: Surviving in a Social Media World of Hackers, Terrorists, Russians, and Fake News, New York, NY: Harper Collins, 2018.

International Workshop on Modeling Social Media (MSM'2019) Chairs' Welcome

Welcome to the *TheWebConf 2019 International Workshop on Modeling Social Media (MSM'2019) - Mining, Modeling and Learning from Social Media*. We are celebrating our 10th edition (one decade) of our workshop. Social networks such as Facebook, Twitter, and LinkedIn generate huge amounts of data that require the application of big data analytics to produce meaningful information to both information consumers and data generators. Machine learning and AI techniques are particularly effective in situations where deep and predictive insights need to be uncovered from such social media data sets that are large, diverse and fast changing.

The workshop aims to focus on how to apply data mining, recommendation, machine learning and AI models, algorithms and systems for analytic and predictive modeling on social media, big data, small data, web and sensor data. We received 15 papers from all around the world covering a broad range of topics, and we accepted 9 papers resulting in a 60% acceptance rate. We evaluated them regarding relevance, quality, and novelty, selecting 5 full papers and 4 short papers. Each paper was reviewed by at least 3 reviewers and then decisions were made from the reviews and the workshop chairs.

The list of the 9 papers that were presented in the half-day workshop were:

- Personalized Degrees: Effects on Link Formation in Dynamic Networks from an Egocentric Perspective (full paper)
- Supervised Lexicon Extraction for Emotion Classification (full paper)
- Learning Influence Probabilities and Modelling Influence Diffusion in Twitter (full paper)
- User Donations in a User Generated Video System (full paper)
- Leveraging Motifs to Model the Temporal Dynamics of Diffusion Networks (full paper)
- Onto Model-based Anomalous Link Pattern Mining on Feature-Rich Social Interaction Networks (short paper)
- CrowdGuard: Characterization and Early Detection of Collective Content Polluters in Online Social Networks (short paper)
- Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule (short paper)
- Sensing Social Media Signals for Cryptocurrency News (short paper)

We thank all those who submitted their papers. We are also grateful to the workshop program committee and reviewers for their timely work in reviewing the submitted contributions. Finally, special thanks are due to the members of the WWW Organizing Committee who made this event possible. We hope you enjoy the workshop, the conference and San Francisco.

Martin Atzmueller
Workshop Chair
Tilburg University

Alvin Chin
Workshop Chair
BMW Group

Christoph Trattner
Workshop Chair
University of Bergen

THE WEB CONFERENCE

International Workshop on Modeling Social Media (MSM'2019) Organization

Workshop Chairs:
Martin Atzmueller *(Tilburg University, The Netherlands)*
Alvin Chin *(BMW Group, USA)*
Christoph Trattner *(University of Bergen, Norway)*

Program Committee:
Alejandro Bellogin *(Universidad Autonoma de Madrid, Spain)*
Javier Luis Canovas Izquierdo *(IN3 – UOC, Spain)*
Michelangelo Ceci *(Universita degli Studi di Bari, Italy)*
Michael Granitzer *(University of Passau, Germany)*
Bin Guo *(Northwestern Polytechnic University, China)*
Eelco Herder *(Radboud University, The Netherlands)*
Geert-Jan Houben *(Delft University of Technology, The Netherlands)*
Wangsu Hu *(BMW Group, USA)*
Ralf Klamma *(RWTH Aachen University, Germany)*
Bart Knijnenburg *(Clemson University, USA)*
Dominik Kowald *(Know-Center, Austria)*
Florian Lemmerich *(RWTH Aachen University, Germany)*
Nico Piatkowski *(TU Dortmund, Germany)*
Haggai Roitman *(IBM Research Haifa, Israel)*
Giancarlo Ruffo *(Universita' di Torino, Italy)*
Marc Smith *(Connected Action Consulting Group, USA)*
Longqi Yang *(Cornell University, USA)*
Arkaitz Zubiaga *(Queen Mary University of London, UK)*

Reviewers:
Angelo Impedovo *(Universita degli Studi di Bari, Italy)*
Emanuel Lacic *(Know-Center, Austria)*
Emanuele Pio Barracchia *(Universita degli Studi di Bari, Italy)*
Tomislav Duricic *(Know-Center, Austria)*
Roberto Corizzo *(Universita degli Studi di Bari, Italy)*

THE WEB CONFERENCE

Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule

Alessia Antelmi
Università degli Studi di Salerno
Fisciano, Italy
aantelmi@unisa.it

Delfina Malandrino
Università degli Studi di Salerno
Fisciano, Italy
dmalandrino@unisa.it

Vittorio Scarano
Università degli Studi di Salerno
Fisciano, Italy
vitsca@unisa.it

ABSTRACT

Online Social Networks (OSNs) represent a fertile field to collect real user data and to explore OSNs user behavior. Recently, two topics are drawing the attention of researchers: the evolution of online social roles and the question of participation inequality. In this work, we bring these two fields together to study and characterize the behavioral evolution of OSNs users according to the quantity and the typology of their social interactions. We found that online participation on the microblogging platform can be categorized into four different *activity levels*. Furthermore, we empirically verified that the 90-9-1 rule of thumb about participation inequality is not an accurate representation of reality. Findings from our analysis reveal that lurkers are less than expected: they are not 9 out of 10 as suggested by Nielsen, but 3 out of 4. This represents a significant result that can give new insights on how users relate with social media and how their use is evolving towards a more active interaction with the new generation of consumers.

CCS CONCEPTS

• **Applied computing** → **Social and behavioural sciences**; • **Human-centered computing** → *Social media.*

KEYWORDS

User behavior; Lurker; Participation Inequality; Role Discovery; Online Social Networks; Data-Driven Analysis

ACM Reference Format:
Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. 2019. Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316705

1 INTRODUCTION

The exponential growth in the use of digital devices and the ubiquitous online access allow Online Social Networks (OSNs) to represent a fertile field to collect real user data and to explore OSNs user behavior [6]. During the last decade, researchers started exploring the *evolution* of the online user activity and their OSNs social roles [10, 12]. Simultaneously, the attention toward the question of *participation inequality* began growing up, especially on the identification and profiling of passive users, aka *lurkers*. According to the 90-9-1 rule[1] about participation inequality, engagement seems not to be equally distributed among users of social media. In 2006, the expert of User Interface Jakob Nielsen introduced this rule of thumb, describing an online community as composed by 90% of lurkers, 9% by users who contribute from time to time, and only 1% by users who account for most contributions. According to this rule, all large-scale, multi-user communities and online social networks that rely on users to contribute content or build services share one property: most users do not participate and lurk in the background. Simply ignoring lurkers in social media analysis could, therefore, lead to misjudgment of overall population interests level [5]. However, due to clear privacy reasons, the identification of lurkers is not a trivial task (for example, we cannot access users' login data) and the definition of the concept of *activity* itself is not straightforward. These considerations lead to our main research question: *How can we capture and characterize OSNs user behavior and its evolution over time according to the level and typology of user interaction in online platforms?*

Our study aims to answer this question by observing the activity pattern of 122,894 Twitter users over a period of 4 months - from August 3rd to December 3rd, 2017 - and by analyzing both their visible actions (e.g., tweets, favorites) and their *hidden* activity (e.g., change of the account's screen name) on the social platform, for a final total of over 36 million actions. It is worthwhile noting that in this work the concept of activity covers all possible actions that could be done on the platform, from posting a tweet to modifying one's own Twitter account biography. The rationale behind this choice is that defining and characterizing lurkers based only on the number of tweets posted during a given time interval could be a limitation on the identification of this class of users. In this sense, our work differs from the current literature. As case study, we chose the microblogging platform Twitter because of the easiness of silent information consumption [5] and the openness of its APIs. The contribution of our work can be summarized as follows:

- Based on the data collected from the Twitter users timeline, we found that users can be identified by four levels of activity (*high*, *medium*, *low*, *no activity*).
- Our findings show that active users exhibit the same behavioral patterns despite the volume of their interactions with the social platform and that they can be described by a set of five roles - namely *Tweeter*, *Quoter*, *Retweeter*, *Replyer*, *Liker*.

The characterization of Twitter users in terms of their activity leads us towards a *corollary question* about the proportion between active and passive users within the social platform: *Does the 90-9-1 rule*

[1] https://www.nngroup.com/articles/participation-inequality

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316705

apply in Twitter? While a lot of effort has been made to study the presence of this pyramid of engagement in online platforms like blogs, forums, wikis (e.g., Wikipedia), and system reviews (e.g., Amazon) [3, 7], systematic studies have not been carried out in social networks yet.

- Our data-driven findings suggest that user engagement in social networks can be captured by our *new* four-levels classification, identifying as lurkers 3 users out of 4 - interestingly, less than the proportion suggested by Nielsen.

Some interesting works about online participation patterns and lurkers can be found in [2, 4, 5, 8, 11].

2 METHODOLOGY

Dataset. To build a dataset as heterogeneous as possible, we focused on a random subset of Twitter users. Due to the absence of Twitter APIs specifically designed to handle this task, we performed the following steps to construct our dataset.

(1) We collected a tweets sample through the Twitter streaming APIs, from which we extracted 241,000 unique users (*starting set*). We could not stop our collecting process at this point because this would have meant working on a biased dataset where all users had posted at least one status update[2].
(2) To reduce the bias introduced in the first step, we randomly selected 200 users from the *starting set*, and we collected all their followers (*followers set*).
(3) Then, we randomly picked 50 from the *followers set*. This set of 50 users will represent the *seed* of our dataset.
(4) For each user in the *seed set*, we randomly chose up to 60 his/her followers, adding them to the final dataset (*extended seed set*) and obtaining around 3,000 users.
(5) For each user in the *extended seed set* we repeated step (4), ending up with around 180,000 users.
(6) Finally, we removed all users with a private account, obtaining 131,301 unique users (*final set*).

We crawled all posts (tweets, retweets, quotes, replies), favorites, and profile snapshots of each user in the *final set* during a 4 months period, from August 3rd to December 3rd, 2017. At the end of this process, we cleaned our dataset from all users marked as spammers by the Twitter Support Team. In addition, we also removed all *churners*, typically identified as those users who are registered to an online platform/service but do not *use* it [14]. It is worthwhile to remind that lurkers *do use* to the online platform, even though they do not generate any content. To deal with this problem, we used a strategy similar to the one described by Gong et al. [5]: if a user does not perform any action during the whole period of observation, then we can consider him/her as a churner, and we definitely remove him/her from the dataset. We consider the following activities as an action: posting a new status update (tweet, retweet, quote or comment), tapping a like, following or unfollowing a new account, changing either the screen name, the description, the location, the profile image or the banner image. At the end of this process, our dataset was made up of **122,894** unique users and over **36 million** activities (posts and favorites).

[2]Twitter does not reveal how the samples are generated and does not even guarantee that the sampling ratios are stable [13].

Clustering Features. We identified two sets of features which capture the platform-specific usage and user activity. These features describe the activity-related behavior considering the time dimension (e.g., the number of activities per week) and the typology of activity done (e.g., number of tweets, likes).

To analyze Twitter users according to their *level* of activity, we represent each user with a unidimensional features vector, containing the following information extracted from the Twitter timeline of the referred user. We will refer to this set of features as **quantity-based** features, where an *activity* can be either a tweet, retweet, quote, reply and favorite.

i. **Q-Number**, total number of activities;
ii. **Q-WeeklyFrequency**, average amount of activities per week;
iii. **Q-AverageTime**, average time in days between two consecutive activities.

To study active Twitter users according to the *typology* of the visible activities done on Twitter, we represent each active user with a unidimensional features vector, containing the following information extracted from the Twitter timeline of the referred user. We will refer to this set of features as **typology-based** features.

i. **T-Tweets**, number of tweets posted;
ii. **T-Quotes**, number of quotes posted;
iii. **T-Retweets**, number of retweets posted;
iv. **T-Replies**, number of replies posted;
v. **T-Favorites**, number of favorites posted.

Workflow. *Characterizing Twitter users according to their level of activity.* Our first goal is to identify user roles based on a *quantitative* view of the user activity through a data-driven approach, without attempting to match behaviors to a pre-defined set of roles. We follow a methodology similar to the one described in the studies of O'Donovan et al. [9] and Yang et al. [14]. We employed K-means as clustering algorithm, whose scalability allowed us to analyze the entire dataset. We used the Euclidean distance as distance measure and the Silhouette analysis to automatically decide the proper number of clusters k. To confirm the significance of the cluster distributions, for each feature and for each pair of clusters we applied the ANOVA and the Tukey's HSD tests, to check whether groups in the same sample statistically differ (ANOVA) and which of them in specific have significance differences (Tukey's test). Before applying the clustering algorithm, we first standardized the feature values by removing the mean and scaling to unit variance. For this analysis, we used the **quantity-based** features. We used the opposite value for the *Q-AverageTime* variable to be semantically consistent with the other two features, *Q-Number* and *Q-WeeklyFrequency*.

Characterizing active Twitter users according to the typology of activities. The workflow just described resulted in 4 clusters, each one representing a user group with different activity levels. We focused on three of these clusters, describing users with *high*, *medium* and *low* degrees of interaction. Our purpose is to identify if active users exhibit the same behavioral patterns in terms of the typology of activity posted despite their quantity. We analyzed each cluster independently, following the same methodology illustrated in the previous paragraph. This time we used the **typology-based** set of features.

Table 1: User groups with different levels of Twitter activity.

Cluster Label	Size	Relative Size	Avg Value *Q-Number*	Avg Value *Q-Weekly Frequency*	Avg Value *Q-Average Time*
H-A	6,378	5.18%	2,282.13	133.84	1.29
M-A	13,661	11.11%	1,125.52	66.11	1.17
L-A	41,470	33.74%	100.27	5.90	2.30
No-A	61,385	49.94%	0.0	0.0	-1.0

3 RESULTS AND DISCUSSION

Twitter Users Characterization. Based on the Silhouette value, we found that 4 was the best number of clusters. Table 1 describes clusters size and average feature values.

· **High-Activity (H-A).** With a relative size of 5%, this is the smallest group found. This role is characterized by the highest average values of both *Q-Number* and *Q-WeeklyFrequency* features. This means that the users in this cluster heavily interact with Twitter, performing 19 actions on the social platform on average per day.

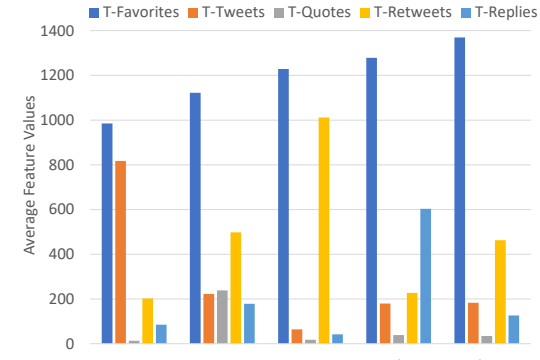
· **Medium-Activity (M-A).** This group is the most similar in terms of behavioral patterns to the *High-Activity* group, albeit with a different magnitude in terms of the overall and weekly activities done. These users interact on a daily basis with Twitter, performing 9 actions on average per day.

· **Low-Activity (L-A)** This cluster is the second larger, with a relative size of 33.74%. These users have an average weekly frequency of almost 6 actions on Twitter, and it is interesting to note that they interact with the social every two days on average.

· **No-Activity (No-A).** Unlike the previous groups, this cluster with a relative size of almost 50% is the largest and most passive group. In fact, as we can observe from Table 1, these users have done no actions over the whole four-months period of observation. The -1.0 value for the *Q-AverageTime* variable means that it was not possible to evaluate this parameter due to the absence of activities. We recall that our dataset does not contain *churners*, thus all users in this cluster have done at least one *hidden* activity.

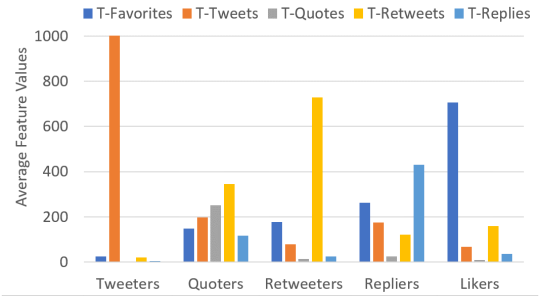
Active Twitter Users Characterization. Based on the Silhouette value, we found that 5 was the best number of clusters for the *High-Activity* and *Medium-Activity* groups; 6 was the best choice for the *Low-Activity* cluster, instead. To verify if the same set of roles occurred in each dataset, we measured the cosine similarity of every pair of role vectors. We used a threshold value equal to 0.75 to determine whether a pair matches. In practice, we found most matching pairs having a cosine similarity grater than 0.9. The sixth role identified in the *Low-Activity* dataset was not associated with any role in the other two groups. Figs. 1, 2 and 3 show the size and average feature values for each cluster. Each cluster approximately corresponds to a role and the role names were selected according to their distinguishing features, in this case, according to the type of the most *common* activity. A description of the characteristics of each cluster role follows.

· **Likers.** This role describes the behavior of the majority of the users (without considering the *Lazy* role identified in the *Low-Activity* dataset). As the label implies, the strictly predominant



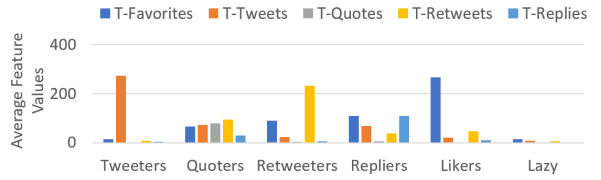
	Tweeters	Quoters	Retweeters	Repliers	Likers
Size	732	496	2,168	801	2,181
Rel. Size	11,47%	7,77%	33,99%	12,55%	34,19%

Figure 1: User roles in the *High-Activity* sub-dataset.



	Tweeters	Quoters	Retweeters	Repliers	Likers
Size	2324	477	2426	681	7753
Rel. Size	17,01%	3,49%	17,75%	4,98%	56,75%

Figure 2: User roles in the *Medium-Activity* sub-dataset.



	Tweeters	Quoters	Retweeters	Repliers	Likers	Lazy
Size	2139	395	2229	1036	5648	30023
Rel. Size	5,15%	0,95%	5,37%	2,49%	13,61%	72,39%

Figure 3: User roles in the *Low-Activity* sub-dataset.

activity done by these users is *tapping a like*. For the *High-Activity* dataset, this typology of activity is a constant across all clusters, suggesting that the huge number of weekly actions done by these users is due precisely to this activity.

· **Retweeters**. Second role in terms of dimension, whose main activity on the microblogging platform is *retweeting*. It is interesting to note how the second most common action in this group is liking and that, specularly, the second most common action done by the *Likers* is retweeting. Liking and retweeting do not produce any new original content; for this reason, we consider them as the most *passive* actions.

· **Repliers.** This group exhibits a behavior similar to the previous ones in terms of *liking* and *retweeting*. However, they mostly mention other users/tweets to express their opinion and share content.
· **Tweeters.** On the contrary, this group tends only to post tweets as its main activity, without interacting with the online platform through other typologies of action. This is especially true for the *Medium-Activity* and *Low-Activity* datasets.
· **Quoters.** In all three sub-datasets, the smallest cluster is represented by this role, whose characterizing activity is the relatively high use of *quotes*, i.e. a retweet plus a personal comment. This group is completely different from the previously described as these users seem to use Twitter homogeneously with regards to the five typologies of activities analyzed.

This second clustering analysis revealed the presence of a conspicuous group in the *Low-Activity* dataset which does not fit in the description of any of the identified roles due to the limited amount of their actions. These users - called **Lazy** - are defined by (i) a low total amount of activities (47.05 on average), (ii) less than 3 activities on average per week (2.76), (iii) *liking* as main activity and (iv) an average time between two consecutive actions of around 4 days. Despite these values, they do not share enough other distinct characteristics for the K-means algorithm to require another additional cluster to differentiate them. Analyzing the first partition obtained using the quantity-based features and a greater value for the parameter k did not directly isolate the *Lazy* group.

Does the 90-9-1 Rule Apply in Twitter? Clustering our dataset according to the *quantity-based* set of features, we ended up with four clusters. If we sort them according to their size (Table 1), we can still identify a pyramid with highly active users at the top and passive users at its bottom. This confirms the general behavior pattern theorized by Nielsen according to whom the majority of the users in an online social media are passive. The deeper analysis performed on the *Low-Activity* cluster highlights the presence of a significant subgroup (24% of the entire dataset) - the *Lazy* cluster - that significantly differs from the other five roles identified. Our interpretation is that the *Lazy* users really belong to the lurker category of *No-Active* users. In Table 2, we report the proportion of our *new* participation inequality hierarchy along with the percentages of contribution - in terms of activities - for each group. Given this consideration, our first important result is that *No-Active users are 75% of the total, thus meaning that lurkers are not 9 out 10 as suggested by Nielsen, but 3 out of 4.* This is a significant result as it can give new insights on how users relate with social media - in this specific case Twitter - and how their use is evolving towards a more active interaction with the new generation of consumers.

A comparison with Nielsen's rule. If one really wants to interpret our results to rewrite the 3-level Nielsen's rule, we can suggest some evidence to favor a 5-20-75 subdivision of the users in *Active*, *Moderately-Active* and *Silent Lurkers* groups, where: the **Active** group corresponds to the *High-Activity* cluster, the **Moderately-Active** group is the result of merging the *Medium-Activity* and the *Low-Activity* (without the subgroup *Lazy*) clusters and the **Silent-Lurkers** group coincides with the *No-Activity* cluster plus the *Lazy* subgroup (Table 3). Despite the fact that our analysis hints toward the 4-level hierarchy described in Table 2, we feel that our *AMS rule*, 5-20-75, may be the correct overall interpretation of our results.

Table 2: The new 4-layers participation inequality hierarchy.

	H-A	M-A	L-A (minus Lazy)	No-A (plus Lazy)
#Users (%)	5.18%	11.11%	9.32%	74.36%
#Actions (%)	40.14%	42.16%	12.31%	5.40%

Table 3: The *AMS* rule for Twitter.

	Active users	Moderately-Active	Silent Lurkers
#Users (%)	5.18%	20.42%	74.36%
#Actions (%)	40.13%	54.47%	5.40%

4 LIMITATIONS AND FUTURE WORK

Further investigation based on real data is needed to justify our "AMS rule", although, in our opinion, a significant step forward has been taken to ground empirical observations (like Nielsen's rule) to the real behavior of users. In this regard, we aim to validate our assumption across more Twitter datasets (e.g., a Twitter community) and domain areas, and verify their distributional results. We plan to enhance our approach - described here and in our previous work [1] - by considering not only the quantitative information but also adding the semantic of the posted content and the structural information derived from the interaction network of the users.

REFERENCES

[1] A. Antelmi, D. Malandrino, and V. Scarano. 2018. Characterizing Twitter Users: What do Samantha Cristoforetti, Barack Obama and Britney Spears Have in Common?. In *IEEE International Conference on Big Data (Big Data)*. 3622–3627.
[2] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. 2009. Characterizing User Behavior in Online Social Networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (IMC '09)*. 49–62.
[3] J. Chan, C. Hayes, and E. Daly. 2010. Decomposing Discussion Forums and Boards Using User Roles. In *ICWSM'10*. 215–218.
[4] R. García-Gavilanes, A. Kaltenbrunner, D. Sáez-Trumper, R. Baeza-Yates, P. Aragón, and D Laniado. 2014. *Who Are My Audiences? A Study of the Evolution of Target Audiences in Microblogs.* 561–572.
[5] W. Gong, E. Lim, and F. Zhu. 2015. Characterizing Silent Users in Social Media Communities. In *ICWSM'15*. 140–149.
[6] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. 2013. Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine* 51, 9 (2013), 144–150.
[7] P. Krammer, M. Kvassay, J. Mojžiš, I. Budinská, L. Hluchý, and M. Jurkovič. 2018. Clustering analysis of online discussion participants. *Procedia Computer Science* 134 (2018), 186 – 195.
[8] M. Muller. 2012. Lurking As Personal Trait or Situational Disposition: Lurking and Contributing in Enterprise Social Media. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. 253–256.
[9] F. T. O'Donovan, C. Fournelle, S. Gaffigan, O. Brdiczka, J. Shen, J. Liu, and K. E. Moore. 2013. Characterizing user behavior and information propagation on a social multimedia network. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW'13)*. 1–6.
[10] M. Revelle, C. Domeniconi, and A. Johri. 2016. Persistent Roles in Online Social Networks. In *Machine Learning and Knowledge Discovery in Databases*. 47–62.
[11] N. Sun, P. PL. Rau, and L. Ma. 2014. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior* 38 (2014), 110 – 117.
[12] O. Varol, E. Ferrara, C. L. Ogan, F. Menczer, and A. Flammini. 2014. Evolution of Online User Behavior During a Social Upheaval. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. 81–90.
[13] Y. Wang, J. Callan, and B. Zheng. 2015. Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API. *ACM Trans. Web* 9, 3 (2015), 1–23.
[14] C. Yang, X. Shi, L. Jie, and J. Han. 2018. I Know You'll Be Back: Interpretable New User Clustering and Churn Prediction on a Mobile Social Application. In *KDD'18*. 914–922.

Personalized Degrees: Effects on Link Formation in Dynamic Networks from an Egocentric Perspective

Makan Arastuie
University of Toledo
Toledo, OH, USA
makan.arastuie@rockets.utoledo.edu

Kevin S. Xu
University of Toledo
Toledo, OH, USA
kevin.xu@utoledo.edu

ABSTRACT

Understanding mechanisms driving link formation in dynamic social networks is a long-standing problem that has implications to understanding social structure as well as link prediction and recommendation. Social networks exhibit a high degree of transitivity, which explains the successes of common neighbor-based methods for link prediction. In this paper, we examine mechanisms behind link formation from the perspective of an ego node. We introduce the notion of *personalized degree* for each neighbor node of the ego, which is the number of other neighbors a particular neighbor is connected to. From empirical analyses on four on-line social network datasets, we find that neighbors with higher personalized degree are *more likely* to lead to new link formations when they serve as common neighbors with other nodes, both in undirected and directed settings. This is complementary to the finding of Adamic and Adar [1] that neighbor nodes with higher (global) degree are *less likely* to lead to new link formations. Furthermore, on directed networks, we find that personalized out-degree has a stronger effect on link formation than personalized in-degree, whereas global in-degree has a stronger effect than global out-degree. We validate our empirical findings through several link recommendation experiments and observe that incorporating both personalized and global degree into link recommendation greatly improves accuracy.

CCS CONCEPTS

• **Information systems** → **Social networks**; *Social networking sites*; *Data mining*; Retrieval models and ranking; • **Human-centered computing** → Social networks.

KEYWORDS

Link formation; egocentric network; personalized degree; dynamic network; link recommendation; link prediction; node degree; common neighbors

ACM Reference Format:
Makan Arastuie and Kevin S. Xu. 2019. Personalized Degrees: Effects on Link Formation in Dynamic Networks from an Egocentric Perspective. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316699

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316699

1 INTRODUCTION

Ever since the rise of different social media platforms, the need for analyzing complex social networks has been growing at an exponential rate. A fundamental problem when analyzing social network data is to understand mechanisms driving link formation. Specifically, in the absence of additional information about the nodes in the network, what properties of the network itself tend to be associated with the formation of future links? Such properties could then be incorporated into models for the evolution of social networks over time [10], which can then be used to predict the formation of links in the future. Predicting future or missing links in a network has been a major area of research and was formalized by Liben-Nowell and Kleinberg [11] as the link prediction problem. It has been widely studied since then, and many different approaches and algorithms have been proposed; we refer readers to Lü and Zhou [13] for a survey of the literature.

General principles that govern the behavior of network dynamics form the core of many link prediction and recommendation problems. Common neighbors [16] and Adamic/Adar [1] are two widely used principles which both predate much of the research on link prediction.

Adamic and Adar [1] found that the "popularity" or global degree[1] of a common neighbor of a pair of nodes has an inverse relationship with the likelihood of the two nodes forming an edge in the future. In this paper, we introduce a new principle that is observed to be present in the dynamics of social networks, the *personalized degree* of common neighbors, and examine its behavior and how it correlates with the formation of future links from an egocentric perspective.

This paper is divided up into three parts. In Section 3, we introduce personalized degree and examine its distribution over several datasets. Section 4 consists of a series of empirical analyses on the effects of personalized degree on future link formation. In Section 5, we validate our empirical findings by incorporating them into two link recommendation algorithms and examining the effects of edge directionality on link recommendation.

Our main contributions are as follows:

- We find that the personalized degree of a neighbor behaves in the opposite manner compared to its global degree; that is, common neighbors with higher personalized degree are *more likely* to be predictive of future links, both in directed and undirected networks.
- In directed networks, we find that personalized out-degree has a stronger effect than personalized in-degree and that

[1] In this paper we will refer to the degree of a node, i.e. the number of nodes a particular node it is connected to, as *global degree* to avoid confusion with personalized degree.

global in-degree has a stronger effect than global out-degree. Consequently, limiting personalized and global degree to personalized out-degree and global in-degree, respectively, provides more accurate predictions of link formations.
- We validate the previous two findings through link recommendation experiments, where we find that incorporating personalized degree alone improves mean link prediction accuracy by 2% to 30% and incorporating both personalized and global degrees improves accuracy by 6% to 35%. Incorporating directions of edges further improves accuracy by up to 11%.

2 BACKGROUND

2.1 Link Prediction

Information in networks is carried out in links between nodes. The presence or absence of a link has implications on the social structure. The importance of links makes the concept of predicting new or missing links in a network extremely desirable. This concept is formalized as the link prediction problem [11] and is widely used in a variety of applications. A few examples include recommendation of people to follow in on-line social media networks [7], imputation of missing links from partially observed networks [9, 13], and validation of models for network formation [10].

All link prediction methods, as mentioned in [11], assign an estimate $\text{score}(x, y)$ to all pairs of nodes (x, y) without any links. All scores are then ranked in decreasing order. The higher the rank, the higher the probability of existing an edge between that pair of node. The earliest and simplest methods for link prediction are based on node neighborhoods, where two nodes x and y are predicted to be more likely to form a link in the future if their sets of neighbors $\Gamma(x)$ and $\Gamma(y)$ have large overlap [11].

The simplest method, typically referred to as just common neighbors (CN), simply uses the number of common neighbors as the predicted score:

$$\text{score}(x, y) := |\Gamma(x) \cap \Gamma(y)|. \tag{1}$$

Another method, commonly referred to as Adamic/Adar (AA), sums over all common neighbors, but weights the common neighbors with a lower global degree more heavily in order to indicate that they are more predictive than the ones with a higher global degree [1]:

$$\text{score}(x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} (\log |\Gamma(z)|)^{-1} \tag{2}$$

The Adamic/Adar link predictor has generally been found to be more accurate than common neighbors on a variety of network datasets [11].

2.2 Egocentric Perspective & Link Recommendation

Link prediction is usually considered at the *global* network level, where the objective is to predict the most likely links to be formed between any pair of nodes in the network that do not already have a link. Moreover, it has also been done at the level of an ego node, where the objective is to predict the most likely links to be formed involving the ego node. This problem is commonly referred to as *link recommendation* [22], because the predicted nodes are often used as recommendations for the ego node, e.g. the Who to Follow feature in Twitter [7].

In this study we take an egocentric perspective. Throughout most of our analyses, we first select a node to be the ego and its neighbors to be the common neighbors. Next, we examine link formation between the ego and all neighbors of the common neighbors that are not common neighbors themselves; or simply put, nodes that are 2 hops away from the ego node. This is specially important since most links are formed between nodes that are 2 hops away due to the locality of link formation [21].

Table 1: Summary statistics of datasets used in this paper.

Dataset	Facebook	Google+	Flickr	Digg
Directed?	No	Yes	Yes	Yes
# of Nodes	63,731	∼29M	∼2.3M	63,740
# of Edges	817,090	∼462M	∼33M	837,104
# of Snapshots	10	4	5	16
Duration (days)	∼869	∼98	∼198	∼1431

2.3 Related Work

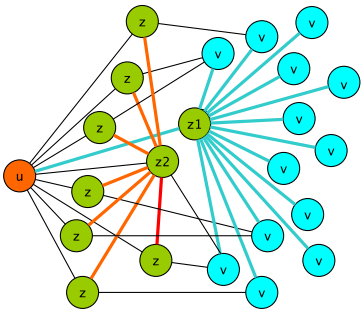
Many link prediction methods have also attempted to incorporate mechanisms behind link formation into link prediction. Leskovec et al. [10] analyze the microscopic evolution of social networks and incorporate their findings on node and link duration as well as triangle closing into a network evolution model that can be used to predict future links. Liu et al. [12] investigate the role of community structure on link formation and propose a link prediction algorithm based on a model for community detection. Cannistraci et al. [5] introduced the concept of local-community-paradigm and suggested that two nodes are more likely to link together if their common neighbors are highly connected within themselves. They referred to the edges between common neighbors as local-community-links (LCL). Wu et al. [20], inspired by LCL and the fact that links between common neighbors of two nodes form a triangle, claimed that the probability of two nodes forming an edge has a direct relationship with the number of triangles passing through their common neighbors and an inverse relationship with their global degree, similar to AA. To the best of our knowledge, these types of studies have usually taken place at the global network level, not from the egocentric perspective that we consider.

2.4 Datasets

In this study, we analyze three directed and one undirected on-line social networks, as shown in Table 1. In all four networks, nodes represent users of that platform.

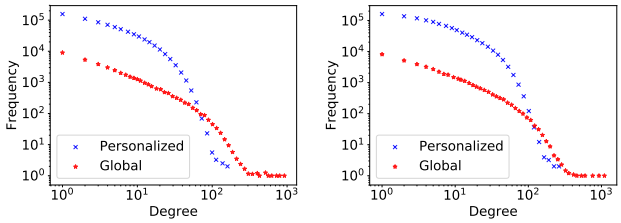
2.4.1 Facebook. The Facebook New Orleans friendship network is an undirected network, collected by Viswanath et al. [18], in which every link represents a friendship between two nodes with a time-stamp of when the friendship was formed. The data trace covers the period between September 2006 and January 2009, and each snapshot lasts for 90 days.

Around March 2008, Facebook introduced the "People You May Know" (PYMK) feature to recommend new friends to their users. While link formations prior to the introduction of PYMK were



Neighbor node z1: Global degree 14, personalized degree 0
Neighbor node z2: Global degree 9, personalized degree 7

Figure 1: Example of a high global and a high personalized degree in a 2-hop egocentric network.



(a) Before PYMK **(b) After PYMK**

Figure 2: Facebook's global and personalized degree distributions before and after the introduction of the People You May Know (PYMK) feature.

likely the result of an organic process, the PYMK feature is likely to influence the link formation process. Thus, we divided the entire Facebook dataset into two sets, before and after PYMK. The first 6 extracted snapshots consist entirely of link formations before PYMK, and the remaining 4 came after PYMK.

2.4.2 Google+. This dataset was crawled from the Google+ network, starting from July 6 until October 11, 2011, by Gong et al. [6], and it covers more than 70% of the entire Google+ network at that time. Incoming and outgoing friends are represented by directed links which belong to one of the 4 predefined snapshots.

Due to the large network size, we randomly selected about 500,000 nodes that appeared in the first snapshot of the network to serve as ego nodes in all of our analyses. The same set of randomly selected nodes are used in all experiments for consistency. Furthermore, if an ego node had more than 100,000 nodes that were at most 2 hops away, then it was not used in the experiments.

2.4.3 Flickr. Flickr is a media hosting service where users can host images and videos and share them with their friends. Most snapshots in the Flickr dataset [14] are 30 days long. Moreover, in order to manage the large size of this network, we made the exact same random sampling decisions as we did for Google+.

2.4.4 Digg. Digg is a news aggregator where users can follow each other, share their own news or blog posts, and up/down vote others' posts. We divided the Digg friendship dataset [8] into 90-day long snapshots, also ignored all self-loops, inactive users, and users with an out-degree of zero.

3 PERSONALIZED DEGREE

For any given ego node, we define the *personalized degree* of a neighbor node to be the number of other nodes that have links with both the ego and the neighbor node. Thus, while each node in a network has a single global degree, it has a *different* personalized degree with respect to every ego node. As shown by node z1 in Figure 1, a neighbor node may have a high global degree, but a low personalized degree. Moreover, personalized degree of a neighbor node is always at least 1 less than its global degree and at least 2 less in order to be a common neighbor between the ego node and a node 2 hops away from the ego.

In directed networks, just like global degree, personalized degree is divided into personalized in-degree and personalized out-degree. These respectively refer to the number of predecessors and successors of the neighbor node that also have an edge with the ego, where the direction of the edge between the ego and all other nodes, including the neighbor node, is the same.

It is important not to confuse personalized degree with localized clustering coefficient, first introduced by Watts and Strogatz [19]. Given a node, localized clustering coefficient measures how close its neighbors are to forming a clique. In measuring local clustering coefficient, we fix one node and look at the number of its formed triangles over the total possible number of triangles with its neighbors. In personalized degree, every neighbor of a neighbor node does form a triangle, however, we not only consider the total number of possible triangles, but we also fix two nodes of each triangle, the ego and the common neighbor, instead of only one fixed node.

3.1 Personalized Degree Distribution

First, we examine both directed and undirected personalized degree distributions to better understand their behavior. We consider each node in the network to be an ego and compute the personalized degree of all of its neighbors. We then plot the distribution of all computed personalized degrees on a log-log scale, similar to plotting the global degree distribution of a network. In addition, in directed networks, in order to stay consistent with our egocentric perspective, we consider all successors of an ego to be its neighbors. Keep in mind that these neighbors may or may not be the predecessors of the ego. Moreover, if it is possible for a node degree to be zero, to observe these instances on a log-log scale, we shift all node degrees up by one. Due to the large number of nodes in all analyzed networks, we plot degree distributions using uniform log-binned scatter plots.

A widely accepted claim in modern network science is that most complex networks, including social networks, are scale-free [3, 4, 15], meaning that their degree distribution follow a power law. As it is shown in Figures 2 and 3, the global degree distribution of both Facebook and Flickr seem to follow a power law as well; nevertheless, Flickr seem to be more strongly scale-free compared to Facebook. Furthermore, both personalized degree distributions seem to follow a power law distribution; however, they both have

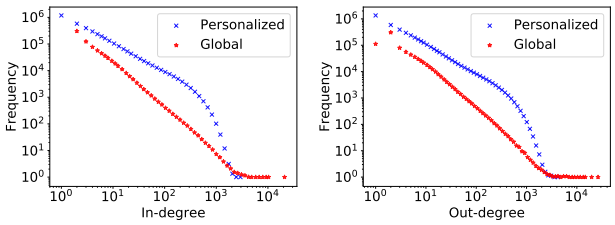


Figure 3: Flickr's global and personalized in-degree and out-degree distributions.

lighter tails than a normal power law. Due to the light-tailed characteristic of these distribution, it is more fitting to describe their distribution by a truncated power law rather than a regular heavy-tailed power law distribution. Degree distributions of both Google+ and Digg are very similar to that of Flickr.

4 EMPIRICAL OBSERVATIONS

In this section, we perform an empirical analysis on the role of common neighbors' degrees, both global and personalized, on link formation from an egocentric perspective on all four datasets.

4.1 Empirical Analysis Procedure

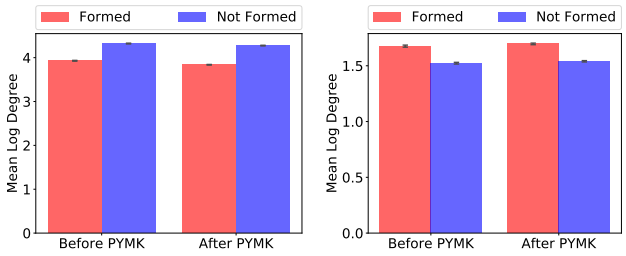
To investigate the role of neighbor node degrees on link formation from an egocentric perspective, we consider every node in the network to be an ego node, and then we analyze the possibility of link formation between the ego node and all nodes which are 2 hops away from it, across all snapshots. More specifically, for every network, beginning with the first snapshot, we consider every node to be an ego node, then for each node v that is 2 hops away from the ego, we check whether it formed an edge with the ego node in the next snapshot. We separate the set of nodes that formed edges with the ego node, denoted by V_f, from the set of nodes that didn't, denoted by V_{nf}.

For each node $v_f \in V_f$, we calculate the mean log global and mean log personalized degrees of all common neighbors between v_f and the ego node. As we observed in Section 3.1, both degrees follow either a truncated or a regular power law distribution, thus simply taking the mean of global and personalized degrees will not result in a reliable summary statistic. Thereby, we took the log of both global and personalized degrees to adjust for this behavior.

Next we take the average of mean log global and mean log personalized degree over all nodes in V_f. We repeat the same process for the nodes $v_{nf} \in V_{nf}$ that did not form an edge with the ego node. Finally, we take the mean over all snapshots of the network.

The aforementioned process is repeated for, and averaged over, all ego nodes. We then compare the mean log global and mean log personalized degrees for all nodes which formed an edge with an ego to the ones that did not.

Note that in order to make sure the averages for both formed and not-formed groups are over the exact same ego nodes in all snapshots, for every ego node, we excluded snapshots where either formed or not-formed group was empty. Therefore, if all snapshots of an ego node were excluded, the ego node was naturally removed from the analysis.



(a) Global degree (b) Personalized degree

Figure 4: Mean log global and mean log personalized degrees (±2 standard errors) of common neighbors in Facebook data before PYMK (19,162 ego nodes) and after PYMK (39,282 ego nodes).

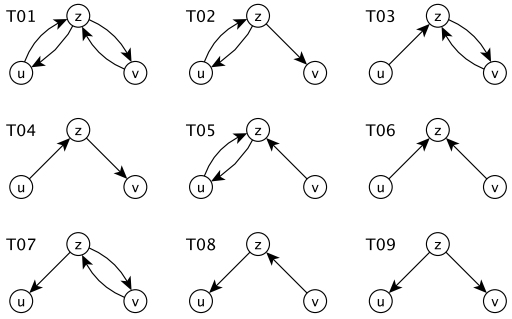


Figure 5: Different open triad patterns in a directed network. Triad labels match those used in [17].

4.2 Observations on Facebook Data

As shown in Figure 4a, we observe that, on average, the global degree of common neighbors of nodes which ended up forming an edge is lower than the ones that did not. Thus, our findings on the Facebook dataset are consistent with those of Adamic and Adar [1] both before and after the introduction of PYMK.

The behavior of personalized degree is shown in Figure 4b. Notice that personalized degree behaves in the exact *opposite* manner compared to global degree. That is, common neighbors of the nodes which formed an edge with the ego node tend to have a *higher* personalized degree, and this is true both before and after PYMK.

4.3 Observations on Directed Networks

In the directed setting, we need to define which nodes we consider as an ego's neighbors and which nodes are 2 hops away. In Figure 5, consider node u to be an ego, node z to be ego's neighbor, and v be a node that is two hops away from the ego. As depicted, there are 9 distinct possible ways that we can select nodes for our empirical analysis and that is without considering whether node v follows the ego (node u). In order to isolate the results for each triad type, we analyze each one separately. Additionally, we also need to consider the direction of edges. The rest of the empirical analysis will be the same as Section 4.1.

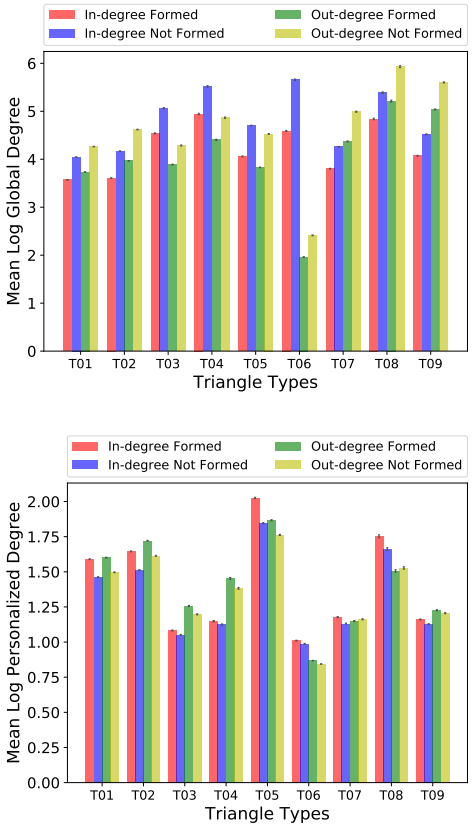
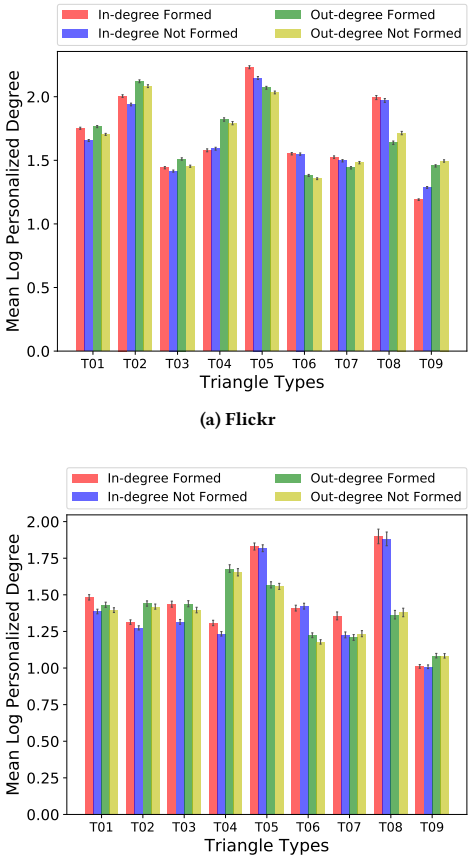


Figure 6: Mean log global and mean log personalized in- and out-degrees (±2 standard errors) of common neighbors in Google+ data (494,881 ego nodes).



(a) Flickr

(b) Digg

Figure 7: Mean log personalized in- and out-degrees (±2 standard errors) of common neighbors in Flickr data (515,624 ego nodes) and Digg data (44,951 ego nodes).

In directed settings our main findings are similar to those on the Facebook network. As shown in Figure 6a, we observe that common neighbors with lower global in-degree or global out-degree are more predictive of future interactions in all 9 different triad types on Google+. This behaviour is also observed in both Flickr and Digg datasets. It is also conveyed from the figure that this observation holds much stronger for some triads than others (e.g. T01 vs. T06). Thereby, we can conclude that the findings of Adamic and Adar [1] also hold in directed networks.

Similar to our findings in Section 4.2, in the majority of triads in all directed datasets, personalized in- and out-degrees behave in an opposite manner compared to global degrees, as shown in Figure 6b for Google+ and Figure 7 for Flickr and Digg. That is, common neighbors with higher personalized in- and out-degrees are more predictive of future links. This holds much stronger for Google+, compared to Digg and Flickr, however the conclusion still holds in the majority of triad types on all three directed data sets. The larger standard errors in Figure 7b are due the smaller number of analyzed nodes in the Digg dataset. The 3 triad types where it seems that this observation does not strongly hold across networks are T07, T08, and T09. In all three cases, neighbors of the ego are considered to be ego's predecessors only. It is uncommon in most link prediction analyses to only consider a node's predecessors as its neighbors.

The main implication of this empirical analysis is that the high degree of a neighbor does not necessary make it less predictive of future links as implied by Adamic and Adar [1]. If the high global degree of a common neighbor is caused by a high personalized degree, then that neighbor is in fact *more predictive* of future links.

5 LINK RECOMMENDATION

Given our egocentric perspective approach concerning link formation, it is only fitting that we try to further validate our findings on personalized degree by incorporating it into link recommendation algorithms. If personalized degree does indeed play a crucial role in link formation, then a link recommendation algorithm that utilizes personalized degree should be more accurate than one that does not.

We first demonstrate a few possible ways to integrate our findings on personalized degree into simple neighborhood-based link recommendation algorithms, namely common neighbors [11] and

Adamic/Adar [1]. Next, we evaluate these personalized degree-based methods on all four datasets. It is important to note that there are many more complex (and likely more accurate) link recommendation algorithms present in the literature as discussed in Section 2.3. The link recommendation methods introduced in this section are quite simple and may not be the most accurate. However, the main purpose of this section is to first validate our empirical findings and consequently demonstrate the possibility for improving node neighborhood-based link recommendation algorithms by taking advantage of personalized degree. We believe that it is possible for more complex algorithms to benefit from personalized degree in a similar manner.

5.1 Proposed Link Recommendation Methods

5.1.1 Personalized Degree Common Neighbors (PD-CN). We first integrate our empirical findings into the common neighbors link prediction method [11]. The common neighbors (CN) link predictor assigns the score defined in (1) to each pair of nodes without a link.

As we observed in Section 4, neighbor nodes of the ego with a higher personalized degree are more predictive of future links. Thus, we should give these neighbor nodes higher weight in the link prediction score. We propose to sum over the log of personalized degrees of common neighbors. We take the log due to the heavy-tailed distribution of personalized degree as discussed in Section 3.1. Hence, the weight of each common neighbor will be the log of its personalized degree with the ego. Weighting common neighbors by degree has been used with global degree in Adamic/Adar, where common neighbors are weighted by the inverse log of their global degree [1].

Therefore, PD-CN for an ego node u and a candidate node v is as follows:

$$\text{PD-CN}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \log\left(|\Gamma(u) \cap \Gamma(z)| + 2\right) \tag{3}$$

where node z is a common neighbor of nodes u and v. The overlap of the neighborhoods of the ego node and the common neighbor correspond to the personalized degree of the common neighbor. We shift personalized degree up by 2, since if it is 0 the log is undefined and if it is 1 then the log is 0, in which case the common neighbor is discarded from the sum.

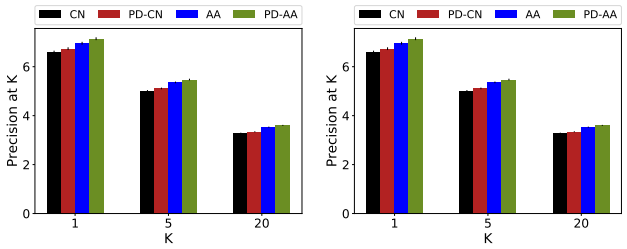
5.1.2 Personalized Degree Adamic/Adar (PD-AA). Here, we integrate personalized degree into the Adamic/Adar (AA) link predictor. Based on our empirical observations, both higher personalized degree and lower global degree of a neighbor node will lead to future link formation; therefore, the ratio of the personalized and global degrees can be a useful measure.

For every common neighbor node z, let $P_z = |\Gamma(u) \cap \Gamma(z)| + 1$ denote its personalized degree shifted up by 1 and $G_z = |\Gamma(z)| + 1$ denote its global degree shifted up by 1. Then, our proposed PD-AA score is given by

$$\text{PD-AA}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \left(\log\left[P_z \frac{(G_z - P_z)}{G_z} + G_z \frac{(G_z - P_z)}{P_z}\right]\right)^{-1} \tag{4}$$

PD-AA score tries to keep the intuition behind AA, that a higher global degree should result in a lower score while also allowing a higher personalized degree to result in a higher score as well. In order to avoid division by zero, we shift personalized degree up by 1 and since it is a subset of global degree we similarly need to shift up the global degree.

Notice that the overall structure of (4) is very similar to the AA score (2). In order to achieve a good balance between global and personalized degrees, we separate the term inside the log into two parts: the weight with respect to P_z and the weight with respect to G_z. We also use a ratio involving the difference between $(G_z - P_z)$. Here, as P_z approaches 1, PD-AA becomes a harsher version of AA with almost 2 times the penalization for global degree, and as P_z approaches G_z, it tries to subtract twice the personalized degree from global degree to decrease the penalty for higher global degree.



(a) Before PYMK (b) After PYMK

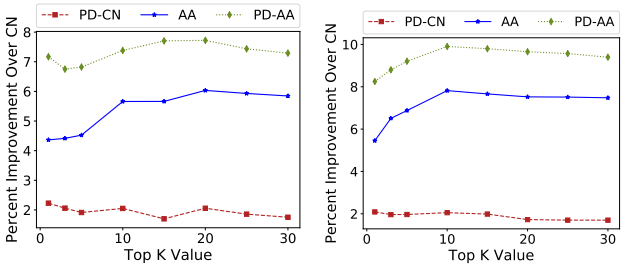
Figure 8: Mean top-K predictive rate of all methods (± standard error) for the Facebook data.

5.2 Evaluation Approach

We test both of our proposed methods, PD-CN and PD-AA, and compare them to CN and AA, respectively. For every snapshot of a network, we compute each of the link prediction scores for all node pairs (u, v) where u is an ego node and v is a node that is 2 hops away from the ego. Next, for each method, we rank all nodes 2 hops away from the ego based on the method's score in deceasing order. Then, we choose top-K predictive rate, also known as precision at K (P@K), to be our evaluation metric. It is the percentage of correctly classified positive samples among the top K instances in the ranking by a specified link predictor. It has been used by others for link recommendation [2, 22], and we view it as the most relevant accuracy metric in a recommendation setting since these are the people who are potentially going to be recommended to a user by a feature such as PYMK in Facebook. Since P@K is heavily dependent on K [21], we chose several K values in all of our evaluations. For each method used in an egocentric network, its top-K predictive rate at a specific K is the mean over all snapshots.

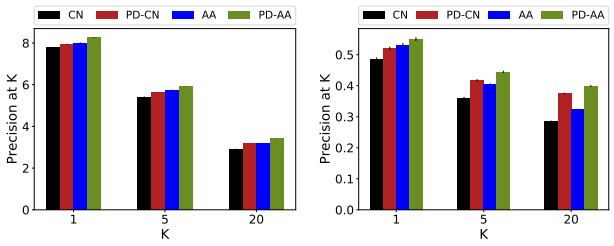
5.3 Results on Facebook Data

The link recommendation results for each of the algorithms on the Facebook dataset, both before and after the introduction of PYMK are shown in Figure 8. Notice that, for each value of K, incorporating personalized degree (PD-CN) consistently improves over CN. However, incorporating global degree (AA) has a greater improvement on link recommendation accuracy. The greatest improvement

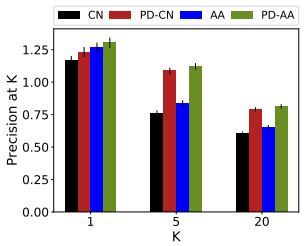


(a) Before PYMK (b) After PYMK

Figure 9: Percent improvements of all link prediction metrics compared to common neighbors as the base metric for Facebook data.

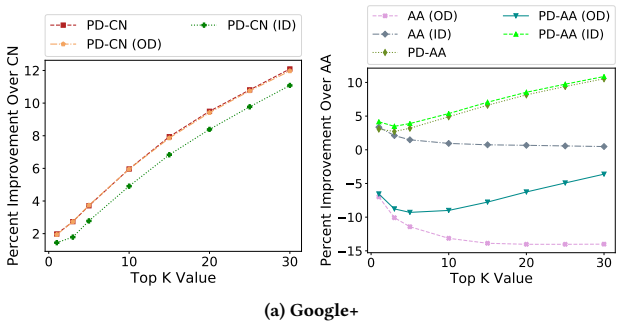


(a) Google+ (b) Flickr

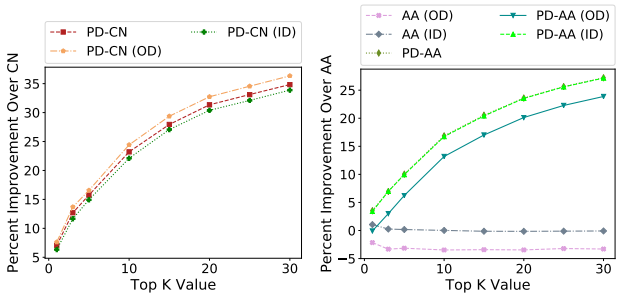


(c) Digg

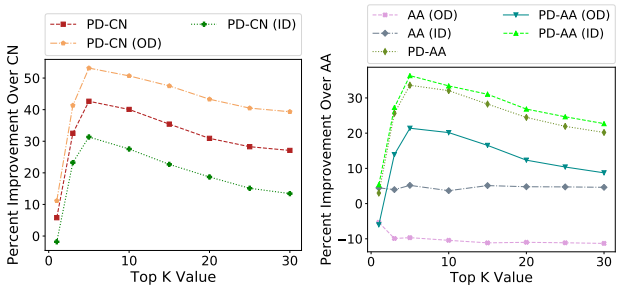
Figure 10: Mean top-K predictive rate of all (undirected) methods (± standard error) for directed datasets.



(a) Google+



(b) Flickr



(c) Digg

Figure 11: Percent improvements of all CN- and AA-based in- and out-degree link prediction metrics, respectively compared to CN and AA as the base metric. *(ID)* and *(OD)* indicate the *in-degree only* and *out-degree only* variants, respectively.

is seen when both personalized and global degrees are included (PD-AA). We also calculate the percentage improvement of all measures with respect to common neighbors, as shown in Figure 9. Here we can also observe the improvements in each method caused by the incorporation of personalized degree. We find that incorporating personalized degrees improves link recommendation accuracy by about 2%, incorporating global degree (AA) improves it by about 4% to 8%, and incorporating both improves it by about 6% to 10%. These improvements further validate our observations regarding the behavior of personalized degree.

5.4 Results on Directed Networks

Similar to our prior directed analyses, we consider all successors of an ego node as its neighbors (z nodes), and all the successors and predecessors of each neighbor node, that are not themselves neighbors of the ego, to be the nodes that are two hops away form the ego (v nodes). Additionally, given that in a directed setting we also have access to the direction of edges, we consider three variants of each method, except for CN since the degree of a node is not a factor. The first variant is turning the network into undirected by reciprocating all edges, the second one is only using in-degree, and lastly only using out-degree. In the case of in- and out-degree variants for AA, we shift up both in- and out-degrees by 2 to avoid taking the log of or dividing by 0. Similarly, in case of PD-AA, since there is no guarantee for global in- or out-degree to be higher than personalized in- and out- degree, respectively, we add 1 to the definition of G_z as described in 4. The evaluation approach is the same as explained in Section 5.2.

The results for all three directed datasets are shown in Figures 10 and 11. We note that link formation in both Flickr and Digg occurs a lot less frequently compared to Google+, and thus, it becomes a much harder task to accurately predict future links, resulting in the much lower values of P@K for Flickr and Digg in Figure 10, but trends are consistent across datasets. In Google+, Flickr, and Digg, PD-CN improved CN by an average of 12%, 23% and 30%, respectively (averaged over values of K). Moreover, only considering global degree (AA) improved prediction accuracy over CN, as well, by 9.4%, 12%, and 9.0% for Google+, Flickr, and Digg, respectively. However, unlike our results in the undirected setting, only considering personalized degree (PD-CN) results in much higher improvements compared to only considering global degree (AA).

Additionally, as shown in Figure 11, although all three variants of PD-CN will result in an improvement over CN, only considering personalized in-degree will decrease P@K compared to the undirected method, while only utilizing personalized out-degree may improve the overall accuracy, sometimes by as much as 10%, as shown in Figure 11c. Similarly, analyzing the effects of in- and out-degree on AA, it is clear that limiting global degree to out-degree will result in a decrease in P@K for all K values, while limiting it to in-degree may result in improvements as it is depicted in Figure 11c. Moreover, while all three variants of PD-AA out-perform all variants of AA in all datasets for all values of K, the impacts of degree directionality on PD-AA are similar to that of AA. Thus, if we consider global degree, personalized degree, as well as the direction of edges, the in-degree variant of PD-AA outperforms all other methods. Comparing it to CN as the base method, average percent improvements for Google+, Flickr, and Digg are 21%, 31% and 37%, respectively. Hence, it is safe to conclude that utilizing global in-degree and personalized out-degree can boost link recommendation accuracy.

6 CONCLUSION

In this paper, we examined the role of node degrees in link formation from an egocentric perspective. We introduced the notion of a personalized node degree for the neighbors of the ego as the number of other neighbors a particular neighbor is connected to. Based on the result of our empirical study on four different datasets, we conclude that the personalized degree of common neighbors behaves in the opposite manner of their global degree when it comes to link formation. While Adamic and Adar [1] found (and we also verified on our datasets) that common neighbors with high global degree are less likely to be predictive of future links, we find that common neighbors with higher personalized degree are *more likely* to be predictive of future links. We also demonstrated an approach to incorporate our empirical findings into node neighborhood-based link recommendation algorithms, and we were able to verify that it leads to improvements in link recommendation accuracy. Furthermore, in the case of directed links, we found personalized out-degree and global in-degree to be more predictive than personalized in-degree and global out-degree, respectively.

We believe our findings have many implications towards understanding link formation from an egocentric perspective. We demonstrated how our findings can be used to improve link recommendation algorithms. Another potential application is towards generative models for network growth. Such models typically operate at a global level, and node degree plays an important role in many models, including the model of Barabási and Albert [3] to generate scale-free networks. Incorporating personalized degree into such models may allow for more accurate replication of link formations in real networks.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation grant IIS-1755824. We would like to thank the University of Toledo's Office of Undergraduate Research for providing funding for Makan Arastuie through the Undergraduate Summer Research and Creative Activities Program (USRCAP).

REFERENCES

[1] Lada A. Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social Networks* 25, 3 (2003), 211–230.
[2] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. 635–644.
[3] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
[4] Albert-László Barabási and Eric Bonabeau. 2003. Scale-free networks. *Scientific American* 288, 5 (2003), 60–69.
[5] Carlo Vittorio Cannistraci, Gregorio Alanis-Lobato, and Timothy Ravasi. 2013. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific Reports* 3 (2013).
[6] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. 2012. Evolution of social-attribute networks: Measurements, modeling, and implications using Google+. In *Proceedings of the ACM Internet Measurement Conference*. 131–144.
[7] Pankaj Gupta, Ashish Goel, Jimmy Lin, and Aneesh Sharma. 2013. WTF: The who to follow service at Twitter. In *Proceedings of the 22nd International Conference on World Wide Web*. 505–514.
[8] Tad Hogg and Kristina Lerman. 2012. Social dynamics of Digg. *EPJ Data Science* 1, 1 (2012), 5.
[9] Gueorgi Kossinets. 2006. Effects of missing data in social networks. *Social Networks* 28, 3 (2006), 247–268.
[10] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. 2008. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 462–470.
[11] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.
[12] Zhen Liu, Jia Lin He, Komal Kapoor, and Jaideep Srivastava. 2013. Correlations between community structure and link formation in complex networks. *PLoS ONE* 8, 9 (2013), e72908.
[13] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A* 390, 6 (2011), 1150–1170.
[14] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2008. Growth of the Flickr social network. In *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks*.
[15] Michael Mitzenmacher. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1, 2 (2004), 226–251.
[16] Mark Newman. 2001. Clustering and preferential attachment in growing networks. *Physical Review E* 64, 2 (2001), 025102.
[17] Daniel Schall. 2015. *Social network-based recommender systems*. Springer.
[18] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. 2009. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*. 37–42.
[19] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 440.
[20] Zhihao Wu, Youfang Lin, Jing Wang, and Steve Gregory. 2016. Link prediction with node clustering coefficient. *Physica A: Statistical Mechanics and its Applications* 452 (2016), 1–8.
[21] Yang Yang, Ryan N. Lichtenwalter, and Nitesh V. Chawla. 2015. Evaluating link prediction methods. *Knowledge and Information Systems* 45, 3 (2015), 751–782. arXiv:1505.04094
[22] Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. 2010. A unified framework for link recommendation using random walks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. 152–159.

Onto Model-based Anomalous Link Pattern Mining on Feature-Rich Social Interaction Networks

Martin Atzmueller
Tilburg University, Department of
Cognitive Science & Artificial Intelligence
Computational Sensemaking Lab
Tilburg, The Netherlands
m.atzmuller@uvt.nl

ABSTRACT

The detection of anomalies and exceptional patterns in social interaction networks is a prominent research direction in data mining and network science. For anomaly detection, typically two questions need to be addressed and defined: (1) What is an anomaly? (2) How do we identify an anomaly? This paper discusses model-based approaches and methods for addressing and formalizing these issues in the context of feature-rich social interaction networks. It provides a categorization of model-based approaches and provides perspectives and first promising directions for its implementation.

CCS CONCEPTS

• **Human-centered computing** → *Social networks*; *Social media*;
• **Information systems** → *World Wide Web.*

KEYWORDS

social media, social network analysis, social interaction networks, feature-rich networks, mining social networks

ACM Reference Format:
Martin Atzmueller. 2019. Onto Model-based Anomalous Link Pattern Mining on Feature-Rich Social Interaction Networks. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316707

1 INTRODUCTION

Social interactions of humans are mediated via social media in various forms and can be modeled using many diverse approaches, particularly using network theory. In the following, we adopt an intuitive definition of social media, regarding it as online systems and services in the ubiquitous web, which create and provide social data generated by human interaction and communication [2]. According to the idea of social interaction networks [2], we focus on interactions between humans, captured by social media. This also includes social relations implemented using specific resources, according to the principle of object-centric sociality [28].

Such data is typically multi-relational, heterogeneous, and usually includes several layers of interdependent temporal abstractions, e. g.,, corresponding to different time intervals of the captured interaction. Then, from a network perspective, these can be conveniently modeled using feature-rich social interaction networks, c. f., [27].

An important and challenging task in such contexts of ubiquitous and social interactions is the detection and analysis of anomalies, e. g., for fraud detection in online social networks, discovering events or unusual topics in heterogeneous network data, or identifying especially interesting or outstanding behavior such as given by influential or "central" actors. From an abstract point of view, an anomaly is defined as a pattern that does not conform to some notion of the expected, normal behavior. Therefore, a straightforward general anomaly detection approach defines a range covering the expected behavior. Then, it identifies any observation in the data that does not belong to this range as an anomaly. This kind of intuitive but relatively simplistic model mainly focuses on point anomalies, as discussed by [1, 43]. However, there is usually no clear formalization of the "normal behavior". Furthermore, current research mostly targets point anomalies, i. e., only relating to individual data points; this does not include anomalies with a more complex structure, e. g., those that encompass a group structure. Therefore, such complex (collective) anomaly patterns are often not detected if the individual contained points seem normal and only their interaction causes an anomaly. In addition, the complexity of anomaly detection is further enhanced by multi-relational and multi-dimensional data, e. g., given by a set of interconnected networks with spatial and temporal characteristics, for example, captured by mobility profiles, or by a set of online social networks and the captured set of activities and/or transactions. Thus, the notion of an anomaly includes other factors compared to a mere outlier which is typically defined by statistical criteria. The concept of an anomaly typically captures more complex criteria, including semantics, (user) expectations and complex data-driven structures.

Moving towards feature-rich social interaction networks, typically two questions then need to be addressed, defined, and formalized in the context of anomalous link patterns:

(1) What is an anomaly (pattern)?
(2) How do we identify an anomaly (pattern)?

One promising option in that context is given by model-based approaches, where an explicit model is used for detecting anomalies. In the following, we outline building blocks of such model-based approaches: We start by targeting the first question in order to address first approaches and promising directions for the second question, which we outline subsequently.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316707

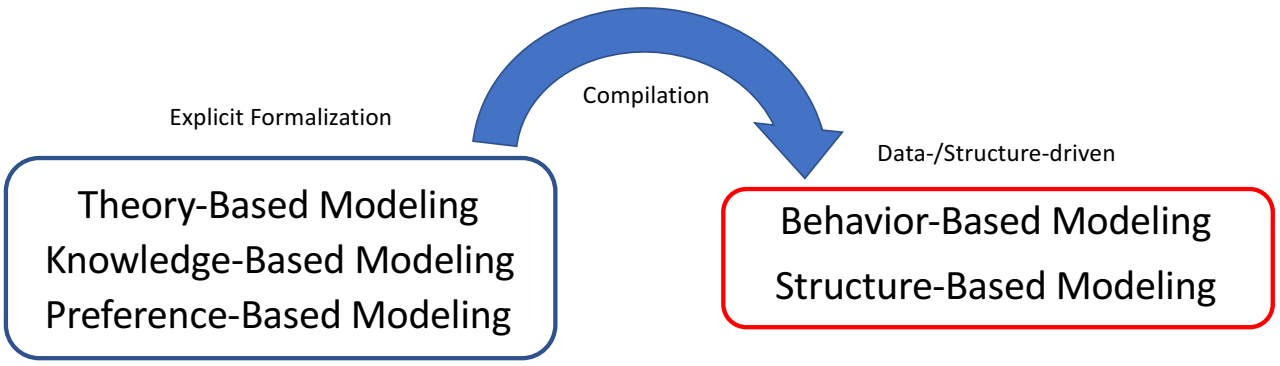


Figure 1: Categorization of Model-Based Approaches for Anomalous Link Pattern Mining

2 OUTLIERS VS. ANOMALIES

There are different definitions of an anomaly. According to the classical definition of [24], "an outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism". Adapted to anomalies in networks (represented by graphs), the general graph anomaly detection problem can be defined as follows: "Given a [...] graph database, find the graph objects [...] that are rare and that differ significantly from the majority of the reference objects in the graph" [1]. Considering networks (represented by graphs), we can focus on different types of graph objects. We can consider individual nodes [1, 46], links between nodes, or more complex substructures of nodes and/or links [6, 14], respectively. Currently, in the literature there are mainly approaches for handling individual point anomalies corresponding to detecting individual nodes, [c. f., 1, 46]. There exist a variety of techniques for anomaly or outlier detection [c. f., 38, 58], e. g., using subspace clustering [e. g., 29, 49], tensor factorization [e. g., 45], or community detection [e. g., 22]. In graph anomaly detection scenarios [c. f., 20, 39, 43] typically static plain graphs are considered.

However, in real-world networks the situation is typically more complex than only considering point anomalies, e. g., with respect to communities, their dynamics, and attribute information assigned to nodes and/or relations. Here, it is also quite difficult to capture the multi-relational nature of inter-connected heterogeneous networks. Therefore, we extend our focus from point anomalies to group anomalies, similar to [57] who define the general group anomaly detection problem as follows: "We are interested in finding the groups which exhibit a pattern that does not conform to the majority of other groups". Furthermore, we consider a range of complex networks, i. e., plain networks, attributed networks, as well as multi-layer networks.

Then, interesting methods for detecting and characterizing anomalous (or exceptional) link structures include exceptional (local) pattern mining and link prediction for anomalous link detection. We briefly discuss both approaches below.

3 MODEL CONSTRUCTION FOR MODEL-BASED ANOMALOUS LINK PATTERN MINING

For anomaly detection on ubiquitous and social interaction networks, we thus focus on the network structures and node properties. Therefore, this involves node topology, node features and attributes of node and/or edges, as formalized in feature-rich networks [27]. Since it is difficult to directly specify what an anomaly is (unless ground truth data is available), we aim for a model-based approach. Then, a specific model identifies a subgraph induced by a set of nodes and/or edges as an anomaly dependent on specific characteristics which are encoded in the respective model. We distinguish between the following options for constructing the models:

(1) Theory-based modeling.
(2) Knowledge-based modeling.
(3) Preference-based modeling.
(4) Behavior-based modeling.
(5) Structure-based modeling.

Regarding that categorization, we basically distinguish between approaches that require an *explicit formalization*, i. e., theory-based, knowledge-based and preference-based modeling, and those that rely on *data-driven* and *structure-driven* criteria, i. e., behavior-based and structure-based modeling. The former can often also be *compiled* into the latter, such that theories or (implicit) knowledge-structures are transformed into explicit structures to be used for anomaly detection (c. f., Figure 1).

3.1 Theory-Based Modeling

As one option for model-based detection of anomalies, we can consider social theories, e. g., homophily. In social interaction networks, homophily [35] has been identified as an important driving factor for establishing contacts, that is actors are more likely to engage with other actors if they are similar with respect to certain attributes, c. f., [10, 36, 37]. Other options include theories like the small world phenomenon [see 53], structural holes [see 18], or the strength of weak ties [see 23]. In [10], we have presented first results on detecting anomaly patterns using a homophilic model. Extensions consider, for example, more refined models on the local pattern formalization, e. g., [4].

3.2 Knowledge-Based Modeling

Knowledge-based approaches for model-based anomalous link pattern mining rely on a form of formalized domain or background knowledge. A powerful representation formalism is given by knowledge graphs, e. g., [16, 25] capturing the relations between concepts, their properties and further (inter-)relations. In such network structures, data is integrated into a comprehensive knowledge model capturing the relations between concepts and their properties in an explicit way, For instance, entities (concepts) are usually represented as nodes, there can be categories (labels) associated to node, and conceptual relations are given by directed edges between the nodes [42]. Following [41], from the point of construction, a knowledge network then mainly describes real world entities and their interrelations. Using such a structure for anomaly detection utilizes the contained relationships and dependencies, comparing expected and deviating paths and relationships on the graph with respect to the observed data. Example applications include, e. g., [8, 15] for anomaly detection in large knowledge graphs compared to complex interaction networks.

3.3 Preference-Based Modeling

Preference-based modeling incorporates a special form of knowledge, i. e., user preferences into building a model of the normal behavior. For example, in interaction networks in the context of social events, e. g., [7] or social interactions on student freshmen weeks [52] preferences can be expressed in order to determine the expected (normal) behavior. Then, e. g., simple approaches for anomaly detection consider correlation-based methods [52] or local pattern mining on the collected interaction data [7] in order to determine normal (expected) or anomalous (deviating) interactions.

3.4 Behavior-Based Modeling

Behavior-based modeling can be considered a data-driven approach, where we consider the respective observed data and compare it to a data-driven reference. This is given, for example, by typical quality function (interestingness functions) in local pattern mining, for example, relating to exceptional model mining and subgroup discovery [3]. Then, observed patterns are compared to the total population, or a null-model of the total population, respectively, as for community detection [6]. Furthermore, we can also analyze compositional structures in social interaction networks, e. g., using compositional subgroup discovery [4]. More complex behavioral modeling approaches consider, for example, Markov chains capturing transition structures in the observed data, in order to detect anomalies [8, 13].

3.5 Structure-Based Modeling

Structure-based modeling approaches take specific (graph-)structures in the network into account, in order to detect specific patterns that conform to those approaches, or to identify deviating groups. Thus, an important focus is the topological structure, for example, considering cliques, hub-authorities, or stars in a network. Then, for local patterns, also descriptive attributes are used for characterizing such structures. Exemplary methods include MinerLC [50] and MinerLSD [14] for exceptional local pattern mining.

4 MINING ANOMALOUS LINK PATTERNS

In the following, we target two directions: First, we consider local pattern mining for identifying exceptional subgroups indicating anomalous link patterns. Second, we focus on link prediction.

4.1 Mining Exceptional Local Patterns

The identification of interesting subgroups (often also called communities) is a prominent research direction in data mining and (social) network analysis, for detecting exceptional local patterns, e. g., [2, 3, 17, 21, 55], e. g., for description, characterization and introspection [5, 6, 9, 11, 12, 14]. Typically, a structural perspective is taken, such that specific (induced) subgraphs are investigated.

Attributed networks, where nodes and/or edges are labeled with additional information, allow further dimensions for detecting patterns that describe a specific subset of nodes of the graph representation of a (social) network As we have outlined above, local pattern mining is a prominent method for mining anomalous link patterns, for theory-based modeling [4, 10], knowledge-based modeling [8, 15], preference-based modeling [7], behavior-based modeling [4, 6] as well as structure-based modeling [14, 50].

4.2 Link Prediction for Anomalous Link Analysis

Link prediction [31, 32] considers the predictive modeling of links between network actors; in social interaction networks, it has a number of prominent applications, e. g., predicting missing links [31], improving collaborative filtering [26], or recommending new contacts [30, 40]. This also relates to mobility [54] and dynamic behavior [51, 56]. First experiments concerning feature-rich networks were presented in [19, 47]. First approaches for analyzing anomalous interrelations were described in [33, 34], also relating to previous work on assessing preferences and actual behavior, as described in [7]. Anomalous link discovery [44] can be implemented using link prediction, essentially focusing on the incorrectly predicted links. While there is no general method fit for all the different anomalous link modeling approaches discussed above, specific link prediction techniques can utilize specific implementations in order to detect anomalies, e. g., [48] for a preference-based approach.

5 CONCLUSION

This paper discussed model-based approaches for anomalous link pattern mining in the context of feature-rich social interaction networks. It categorized those, providing different perspectives, and outlined examples for its implementation. While we observe, that the outlined categories for model-based anomalous link pattern mining are not necessarily mutually exclusive, e. g., considering behavior-based and structure-based approaches, we nevertheless observe the capacity of the discussed approaches. In particular, adapting and extending those from descriptive methods, e. g., using local (exceptional) pattern mining, towards predictive methods for link analysis holds considerable potential for further research.

ACKNOWLEDGMENTS

This work has been partially supported by the German Research Foundation (DFG) project “MODUS” (under grant AT 88/4-1).

REFERENCES

[1] L Akoglu, H Tong, and D Koutra. 2015. Graph Based Anomaly Detection and Description. *Data Min. Knowl. Disc.* 29, 3 (2015), 63.

[2] Martin Atzmueller. 2014. Data Mining on Social Interaction Networks. *Journal of Data Mining and Digital Humanities* 1 (June 2014).

[3] Martin Atzmueller. 2015. Subgroup Discovery. *WIREs DMKD* 5, 1 (2015), 35–49.

[4] Martin Atzmueller. 2018. Compositional Subgroup Discovery on Attributed Social Interaction Networks. In *Proc. International Conference on Discovery Science*. Springer, Berlin/Heidelberg, Germany.

[5] Martin Atzmueller, Joachim Baumeister, and Frank Puppe. 2006. Introspective Subgroup Analysis for Interactive Knowledge Refinement. In *Proc. FLAIRS*. AAAI Press, Palo Alto, CA, USA, 402–407.

[6] Martin Atzmueller, Stephan Doerfel, and Folke Mitzlaff. 2016. Description-Oriented Community Detection using Exhaustive Subgroup Discovery. *Information Sciences* 329 (2016), 965–984.

[7] Martin Atzmueller, Tom Hanika, Gerd Stumme, Richard Schaller, and Bernd Ludwig. 2016. Social Event Network Analysis: Structure, Preferences, and Reality. In *Proc. IEEE/ACM ASONAM*. IEEE Press, Boston, MA, USA.

[8] Martin Atzmueller, Benjamin Kloepper, Hassan Al Mawla, Benjamin Jäschke, Martin Hollender, Markus Graube, David Arnu, Andreas Schmidt, Sebastian Heinze, Lukas Schorer, Andreas Kroll, Gerd Stumme, and Leon Urbas. 2016. Big Data Analytics for Proactive Industrial Decision Support. *atp edition* 58, 9 (2016).

[9] Martin Atzmueller and Florian Lemmerich. 2013. Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. *International Journal of Web Science* 2, 1/2 (2013).

[10] Martin Atzmueller and Florian Lemmerich. 2018. Homophily at Academic Conferences. In *Proc. WWW 2018 (Companion)*. IW3C2 / ACM.

[11] Martin Atzmueller, Florian Lemmerich, Beate Krause, and Andreas Hotho. 2009. Who are the Spammers? Understandable Local Patterns for Concept Description. In *Proc. 7th Conference on Computer Methods and Systems*. Krakow, Poland.

[12] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. 2005. Profiling Examiners using Intelligent Subgroup Mining. In *Proc. IDAMAP*. Aberdeen, Scotland, 46–51.

[13] Martin Atzmueller, Andreas Schmidt, Benjamin Kloepper, and David Arnu. 2017. HypGraphs: An Approach for Analysis and Assessment of Graph-Based and Sequential Hypotheses. In *Postproc. NFMCP (LNAI)*, Vol. 10312.

[14] Martin Atzmueller, Henry Soldano, Guillaume Santini, and Dominique Bouthinon. 2018. MinerLSD: Efficient Local Pattern Mining on Attributed Graphs. In *Proc. 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*.

[15] Martin Atzmueller and Eric Sternberg. 2017. Mixed-Initiative Feature Engineering Using Knowledge Graphs. In *Proc. K-CAP*. ACM Press, New York, NY, USA.

[16] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia-A crystallization point for the Web of Data. *Web Semantics* 7, 3 (2009), 154–165.

[17] Ronald S Burt. 1978. Cohesion Versus Structural Equivalence as a Basis for Network Subgroups. *Sociological Methods & Research* 7, 2 (1978), 189–212.

[18] R S Burt. 1992. *Structural Holes - The Social Structure of Competition*. Harvard University Press, Cambridge, Massachusetts, USA.

[19] Lukas Eberhard, Christoph Trattner, and Martin Atzmueller. 2018. Predicting Trading Interactions in an Online Marketplace through Location-Based and Online Social Networks. *Information Retrieval Journal* (2018).

[20] William Eberle and Lawrence B. Holder. 2007. Anomaly Detection in Data Represented as Graphs. *Intelligent Data Analysis* 11, 6 (2007), 663–689.

[21] Ove Frank. 1997. Composition and Structure of Social Networks. *Mathématiques et sciences humaines. Mathematics and Social Sciences* 137 (1997).

[22] Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. 2010. On Community Outliers and Their Efficient Detection in Information Networks. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. ACM, New York, NY, USA, 813–822.

[23] M S Granovetter. 1973. The Strength of Weak Ties. *The American Journal of Sociology* 78, 6 (1973).

[24] D Hawkins. 1980. *Identification of Outliers*. Chapman and Hall, London, UK.

[25] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194 (2013), 28–61.

[26] Zan Huang, Xin Li, and Hsinchun Chen. 2005. Link Prediction Approach to Collaborative Filtering. In *Proc. JCDL*. ACM, New York, NY, USA, 141–142.

[27] Roberto Interdonato, Martin Atzmueller, Sabrina Gaito, Rushed Kanawati, Christine Largeron, and Alessandra Sala. 2019. Feature-Rich Networks: Going Beyond Complex Network Topologies. *Applied Network Science* ((In Press) 2019).

[28] Karin Knorr-Cetina. 1997. Sociality with Objects: Social Relations in Postsocial Knowledge Societies. *Theory, Culture and Society* 14, 4 (1997), 1–43.

[29] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2012. Subspace Clustering. *WIREs: Data Mining and Knowledge Discovery* 2, 4 (2012), 351–364.

[30] Xin Li and Hsinchun Chen. 2009. Recommendation as Link Prediction: A Graph Kernel-based Machine Learning Approach. In *Proc. JCDL*. ACM, New York, NY, USA, 213–216.

[31] David Liben-Nowell and Jon M. Kleinberg. 2003. The Link Prediction Problem for Social Networks. In *Proc. CIKM*. ACM, New York, NY, USA, 556–559.

[32] L. Lü and T. Zhou. 2010. Link Prediction in Weighted Networks: The Role of Weak Ties. *EPL (Europhysics Letters)* 89 (2010), 18001.

[33] Spyroula Masiala and Martin Atzmueller. 2018. First Perspectives on Explanation in Complex Network Analysis. In *Proc. BNAIC*. Jheronimus Academy of Data Science, Den Bosch, The Netherlands.

[34] Spyroula Masiala and Martin Atzmueller. 2018. Towards Explainable Complex Network Analysis. In *Proc. Dutch-Belgian Database Day*. Hasselt University, Belgium.

[35] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Ann. Rev. Sociology* 27, 1 (2001), 415–444.

[36] Folke Mitzlaff, Martin Atzmueller, Andreas Hotho, and Gerd Stumme. 2014. The Social Distributional Hypothesis. *Journal of Social Network Analysis and Mining* 4, 216 (2014), 1–14.

[37] Folke Mitzlaff, Martin Atzmueller, Gerd Stumme, and Andreas Hotho. 2013. Semantics of User Interaction in Social Media. In *Complex Networks IV*. Studies in Computational Intelligence, Vol. 476. Springer, Heidelberg, Germany.

[38] Emmanuel Müller, Matthias Schiffer, Patrick Gerwert, Matthias Hannen, Timm Jansen, and Thomas Seidl. 2010. SOREX: Subspace Outlier Ranking Exploration Toolkit. In *Proc. ECML PKDD*. Springer, Heidelberg, Germany, 607–610.

[39] Caleb C. Noble and Diane J. Cook. 2003. Graph-Based Anomaly Detection. In *Proc. ACM SIGKDD*. ACM, New York, NY, USA, 631–636.

[40] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. 2011. Friendlink: Link Prediction in Social Networks via Bounded Local Path Traversal. In *Proc. CASoN*. IEEE, Fremont, CA, USA, 66–71.

[41] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web* 8, 3 (2017), 489–508.

[42] Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge graph identification. In *International Semantic Web Conference*. Springer, 542–557.

[43] Stephen Ranshous, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F. Samatova. 2015. Anomaly Detection in Dynamic Networks: A Survey. *WIREs: Computational Stat.* 7, 3 (2015), 223–247.

[44] Matthew J Rattigan and David Jensen. 2005. The case for anomalous link discovery. *Acm Sigkdd Explorations Newsletter* 7, 2 (2005), 41–47.

[45] Anna Sapienza, André Panisson, Joseph Wu, Laetitia Gauvin, and Ciro Cattuto. 2015. Anomaly Detection in Temporal Graph Data: An Iterative Tensor Decomposition and Masking Approach. In *Proc. International Workshop on Advanced Analytics and Learning on Temporal Data (CEUR)*, Vol. 1425. CEUR-WS.org.

[46] David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, and Qingmai Wang. 2014. Anomaly Detection in Online Social Networks. *Social Networks* 39 (2014), 62–70.

[47] Christoph Scholz, Martin Atzmueller, Alain Barrat, Ciro Cattuto, and Gerd Stumme. 2013. New Insights and Methods For Predicting Face-To-Face Contacts. In *Proc. AAAI ICWSM*. AAAI Press, Palo Alto, CA, USA.

[48] Christoph Scholz, Jens Illig, Martin Atzmueller, and Gerd Stumme. 2014. On the Predictability of Talk Attendance at Academic Conferences. In *Proc. 25th ACM Conference on Hypertext and Social Media*. ACM Press, New York, NY, USA.

[49] Thomas Seidl, Emmanuel Müller, Ira Assent, and Uwe Steinhausen. 2009. Outlier Detection and Ranking based on Subspace Clustering. In *Uncertainty Management in Information Systems (Dagstuhl Seminar Proceedings)*, Christoph Koch, Birgitta König-Ries, Volker Markl, and Maurice van Keulen (Eds.). Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, Dagstuhl, Germany.

[50] Henry Soldano, Guillaume Santini, Dominique Bouthinon, and Emmanuel Lazega. 2017. Hub-Authority Cores and Attributed Directed Network Mining. In *Proc. ICTAI*. IEEE, Boston, MA, USA.

[51] Chayant Tantipathananandh and Tanya Y. Berger-Wolf. 2009. Constant-Factor Approximation Algorithms for Identifying Dynamic Communities. In *Proc. ACM SIGKDD* (2009-07-02). ACM, New York, NY, USA, 827–836.

[52] Lisa Thiele, Martin Atzmueller, Simone Kauffeld, and Gerd Stumme. 2014. Subjective versus Objective Captured Social Networks: Comparing Standard Self-Report Questionnaire Data with Observational RFID Technology Data. In *Proc. Measuring Behavior*. Wageningen, The Netherlands.

[53] J Travers and S Milgram. 1969. An Experimental Study of the Small World Problem. *Sociometry* 32 (1969), 425–443.

[54] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. 2011. Human Mobility, Social Ties, and Link Prediction. In *Proc. ACM SIGKDD*. ACM, New York, NY, USA, 1100–1108.

[55] Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications* (1 ed.). Cambridge University Press, Cambridge, MA, USA.

[56] Duncan J. Watts and Steven H. Strogatz. 1998. Collective Dynamics of 'small-world' Networks. *Nature* 393, 6684 (june 1998), 440–442.

[57] Rose Yu, Xinran He, and Yan Liu. 2014. GLAD: Group Anomaly Detection in Social Media Analysis. In *Proc. ACM SIGKDD*. ACM, New York, NY, USA, 372–381.

[58] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. 2012. A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5, 5 (2012), 363–387.

Sensing Social Media Signals for Cryptocurrency News

Johannes Beck*
ETH Zurich
Switzerland
beckjoh@ethz.ch

Roberta Huang*
ETH Zurich
Switzerland
huangr@ethz.ch

David Lindner*
ETH Zurich
Switzerland
lindnerd@ethz.ch

Tian Guo
Computational Social Science,
ETH Zurich
Switzerland
tian.guo@gess.ethz.ch

Ce Zhang
Department of Computer Science,
ETH Zurich
Switzerland
ce.zhang@inf.ethz.ch

Dirk Helbing
Computational Social Science,
ETH Zurich
Switzerland
dirk.helbing@gess.ethz.ch

Nino Antulov-Fantulin
Computational Social Science,
ETH Zurich
Switzerland
anino@ethz.ch

ABSTRACT

The ability to track and monitor relevant and important news in real-time is of crucial interest in multiple industrial sectors. In this work, we focus on cryptocurrency news, which recently became of emerging interest to the general and financial audience. In order to track popular news in real-time, we (i) match news from the web with tweets from social media, (ii) track their intraday tweet activity and (iii) explore different machine learning models for predicting the number of article mentions on Twitter after its publication. We compare several machine learning models, such as linear extrapolation, linear and random forest autoregressive models, and a sequence-to-sequence neural network.

CCS CONCEPTS

• **Networks** → *Social media networks*; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

Social media, mining and learning, cryptocurrency

ACM Reference Format:
Johannes Beck, Roberta Huang, David Lindner, Tian Guo, Ce Zhang, Dirk Helbing, and Nino Antulov-Fantulin. 2019. Sensing Social Media Signals for Cryptocurrency News. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316706

*Authors contributed equally to this research.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316706

1 INTRODUCTION

Social media and news play an important role in driving the fluctuation of economic indicators and financial markets [5, 10, 13, 18] in a nontrivial fashion. Recently, novel financial markets have emerged, that are exchanging between fiat money and cryptocurrencies [9]. As of December 2018, cryptocurrencies have a total market capitalization of $120 billion, with more than 250000 transactions per day. Therefore, it is not surprising that the rapid development of cryptocurrency has attracted increasing attention from news and social media.

A large volume of news articles about cryptocurrencies, published daily can make it hard for individuals or traders to filter out relevant information and make informed decisions in this domain. Fortunately, people share and discuss news every day in large quantities on social media platforms, e.g. on Twitter, which is the focus of this paper. Therefore, social media can be a good proxy to monitor and track "important" news about cryptocurrencies. Our work is motivated by the hypothesis that high engagement with a news article on Twitter is related to the "importance" of an article.

In this paper, we introduce an online data mining system which connects news and tweets discussing it. We also perform preliminary data exploratory and predictive analytics using machine learning and deep learning. Overall, the contribution of this paper is as follows: (i) We build an online data mining pipeline to extract news articles from a discussion on Twitter and collect tweets associated with the articles. This paired news and tweet data is continuously updated in a cloud database. This data is a rich source for studying public interest and attention on cryptocurrency and the potential effect of social media on the market. (ii) Based on the news and associated tweets collected by the pipeline, we perform exploratory data analysis to characterize news discussion on Twitter. (iii) We apply machine learning and deep learning models to predict the popularity of news articles on Twitter. We aim to predict the number of tweets mentioning a news article related to cryptocurrencies, which we consider as a measure of its "importance".

2 RELATED WORK

Many studies have focused on the relationship between social media, news, and other information from the www onto financial markets [1, 6, 13]. However, the main focus of our work is modeling and prediction of news popularity via social media. In [17], the authors link a given news article to social media utterances that implicitly reference it through a dedicated query model. Tracking and automatically connecting news articles to Twitter conversations by Twitter hashtags was studied in [14]. In [3], the authors constructed a multi-dimensional feature space derived from an article and use a conventional SVM to predict its popularity. The authors in [11] show how the class of temporal point processes (Hawkes) can be used for predicting Retweet dynamics. The authors in [8] propose how to leverage knowledge base information for improving popularity prediction. Starting from the idea that only a small amount of news articles become popular, [12] focused on the subset of the most popular news to rank articles. In [2] it formulates article importance prediction as a classification task using SVM.

In this paper, we exploit ensemble machine learning and sequence to sequence (seq2seq) deep learning to study the predictability of crypto news popularity on Twitter in real-time mode. In contrast to others, our analysis is focused on the intraday importance prediction.

3 DATA PIPELINE

The data pipeline consists of a real-time online system, with the following components: (i) Twitter collection, (ii) news article collection, and (iii) tweet-article matching.

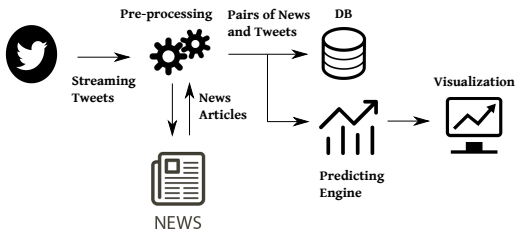


Figure 1: Architecture of the data pipeline.

The Twitter data collection was implemented by using the publicly accessible Twitter streaming API with real-time filtering by a list of cryptocurrency related keywords.

The article data collection is obtained by scraping news from the dynamic set of gazetteer source URLs. The set of gazetteer source URLs is automatically updated by extracting the URLs from the content of downloaded tweets.

The tweet-article matching data is the document-oriented database, that contains matchings between news and tweets. The matching exists if the tweet explicitly contains the URL of an article.

First post-processing step is **merging** of articles. We merge the matchings of two articles if they fulfill the following 3 conditions: (1) the URLs of both articles share the same host as well as the same path; (2) both articles have the same title; and (3) both articles were published at the same time. These conditions allow for merging of same articles of which the URLs have different query strings. While merging the articles we also remove duplicate entries for the same tweets which are sometimes present in the database. For instance, Fig. 2 and 3 demonstrate the top publisher, keywords and news ranking snapshot from the pipeline.

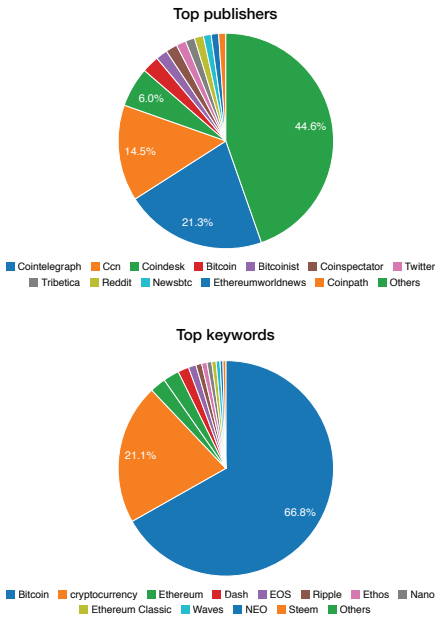


Figure 2: Top publishers and top keywords on 24th Jan 2019, visualized by our system.

#	Article	Publisher	Publication Date	Current	Prediction
1	$20 Million Funding Round in Blockchain Firm Symbiont Includes Citigroup and Nasdaq	Cointelegraph	2019-01-24 01:49	138	166.07
2	Samsung Galaxy S10 Bitcoin Wallet Leaked by Insider: Is it Official?	Ccn	2019-01-24 00:04	147	165.69
3	US: New Hampshire Bill Aims to Legalize Bitcoin for State Payments in 2020	Cointelegraph	2019-01-24 08:30	121	164.10
4	CBOE Withdraws Rule Change Request to List Bitcoin Exchange-Traded Fund	Cointelegraph	2019-01-23 21:20	131	138.92
5	Blockchain Startup to Boost Patient Safety and Prevent Overprescribing	Cointelegraph	2019-01-23 18:00	127	129.03
6	Binance Follows Major Cryptocurrency Exchanges With Launch of OTC Trading Desk	Cointelegraph	2019-01-24 13:40	84	127.02
7	UK Standards Institution Partners With Blockchain Startup for Supply Chain Compliance	Cointelegraph	2019-01-24 12:18	76	115.20
8	The Simple Reasons Why the Bitcoin Price Will Never Go to Zero	Ccn	2019-01-23 22:15	99	106.78

Figure 3: Top article ranked by our system on 24th Jan 2019.

We work with three entities in our data: news articles, tweets, and matchings. Article entity has the following properties: `URL`, `title`, `publication time` and `text`. Tweet entity has the following properties: `user-id`, `text`, `publication time` and `links`. Each matching entity has `article-id` and `list of matched tweets`.

4 PREDICTIVE ANALYTICS

Let t_0, t_1 be two timestamps with $t_0 < t_1 < t_0 + \delta$, where we set δ to 24 hours in this paper. Given an article published at time t_0 and all tweets published between t_0 and t_1 that mention the article, the task is to predict the cumulative number of tweets mentioning the article between time t_0 and $t_0 + \delta$. Here t_1 is the prediction starting time, i.e. how much historical data can be used to predict.

4.1 Feature Extraction

The **time series feature** f_k is given by the number of mentions of the article between t_0 and $t_0 + kd$ where $k \in \{j \in \{1, 2, ...\} | t_0 + jd \leq t_1\}$. As an example suppose that t_1 is 3 hours after t_0 and the article is mentioned twice, once and three times in hours 1, 2 and 3 since publication respectively. Then there are 3 time series features: $f_1 = 2, f_2 = 3, f_3 = 6$. Note that the number of these features is not constant, but depends on t_1.

We extract a vector of **content features** from each article, by using a keyword list to allow the models to learn individual dynamics for articles related to different cryptocurrencies. Each cryptocurrency is represented by a binary feature, that is set to 1 if one of the keywords related to the concept is present in the title of the article.

The amount of Twitter mentions might further be related to the publisher of an article. We then extract the 10 publishers with the highest numbers of mentions. For each of these publishers, we introduce a **context binary feature** set to 1 if the article was published by the respective publisher.

4.2 Predictive Models

As a **baseline model**, we use a linear extrapolation of the last k time series features by fitting a linear function of the time step to the dataset given by $\{(K - k + 1, f_{K-k+1}), ..., (K, f_K)\}$. The model ignores content and context features. In our experiments, we will choose $k = 3$.

An **autoregressive model** (AR) [15] of order k predicts the value at the next timestep $K + 1$ based on the values observed at the previous k timesteps $K - k + 1, ..., K$. In our experiments, we provide K as an additional input to the model. The idea is, that the dynamics can be very different a few hours after the publication and shortly before the end of the prediction window. In our case, we have to predict multiple steps in the future. This is achieved by recursively predicting next timestep and then using it as an input feature for multistep prediction. We use two autoregressive models: (i) linear AR model and (ii) random forest.

A **random forest** [4] is an ensemble of decision trees. The total response of the random forest model is the average prediction of all decision trees. In order to increase the variety of the individual decision trees, each tree is trained on a bootstrapped sample from the original dataset and uses only random subsets of the features for each decision.

Sequence-to-sequence model [16] consists of two recurrent neural networks (RNN): encoder and decoder. The encoder receives as inputs all available time series features. The initial hidden state is given by the final hidden state of the encoder. In our architecture, the output of the decoder at each timestep serves as input for a fully connected network with one hidden layer that outputs the predicted value for the next timestep. If context or content features are used, those features serve as additional inputs to the fully connected network. The predicted value is then used as the input at the next timestep. As a loss function, we use the sum of squared errors between the predicted values and the real values. Our sequence-to-sequence model implementation is based on the gated recurrent unit (GRU) [7], a variant of RNN.

5 EVALUATION AND RESULTS

5.1 Dataset

In this paper, we have done the evaluation of our real-time model with 23535 articles published between 2018-12-02 00:00:00 and 2018-12-09 00:00:00, and all tweets mentioning those articles. We want to estimate confidence intervals of the performance of the different models instead of just obtaining point estimates. Therefore, we use bootstrapping to generate 100 new datasets consisting of 2000 samples from the validation set.

5.2 Set-up

For the baseline model, we choose a linear interpolation of the most recent 3 time series features. The linear autoregressive model is evaluated for orders 1, 3 and 5. The random forest autoregressive models are all of order 3 with 50 and 500 estimators. The sequence-to-sequence model has a hidden state size of 200 for the encoder and the decoder. The hidden dense layer consists of 200 units. The network is then trained in an end-to-end fashion, using back-propagation with training batches of size 64. The baseline model is only provided the time series features. The autoregressive models and the sequence-to-sequence model are trained using the time series, content and context features.

5.3 Evaluation metrics

The goal of our prediction is to extract the most popular cryptocurrency related articles. Therefore, our evaluation focuses on the top k articles with most mentions on Twitter. We use standard *mean absolute percentage error* (MAPE) to measure the accuracy of the predicted number of mentions and *normalized discounted cumulative gain* (NDCG) to measure the quality of the induced ranking of articles. MAPE computes by how many percents the predicted value deviates from the actual value on average, while NDCG value of 1 indicates the correct ranking.

We vary the prediction start time to be 5, 10, 15 or 20 hours after publication time while keeping the target prediction time fixed at $\delta = 24$ hours after the publication.

5.4 Overall performance

In Fig. 4 we see that all models make better predictions, the closer the prediction start time is to the target time. After 15 and 20 hours from the publication time, even the baseline model already achieves very good performance with MAPE of 20%. At prediction start times of 15 and 20 hours, the random forest (RF) and sequence-to-sequence (S2S) model achieve a lower MAPE than the baseline.

At prediction start time 5 hours after publication advanced models achieve a significantly lower MAPE than the baseline. RF and S2S model achieve a MAPE around $30 - 40\%$, while the linear model is at about 45% and the baseline at 70%. For predictions starting 10

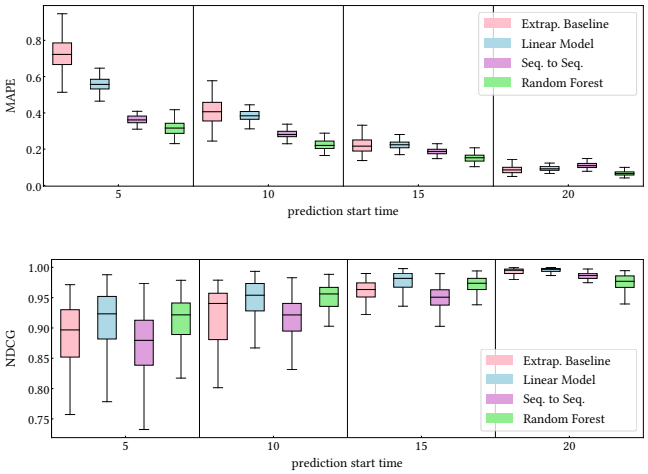


Figure 4: MAPE and NDCG of the different models evaluated on the test data set. The quantiles are determined using 100 bootstrap samples. Predictions are evaluated at 4 different prediction times, 5, 10, 15, and 20 hours after publication of an article.

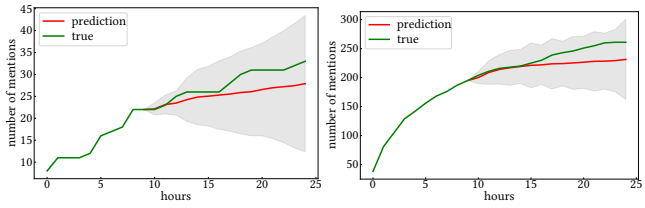


Figure 5: Predictions of the RFAR model for two articles from the test set. The shaded area shows 95% prediction intervals that are determined from the distribution of the predictions by the ensemble estimators.

hours after publication, the baseline and the linear model improve significantly over their performance 5 hours after publication. However, RF and S2S model still achieve a significantly lower MAPE. It is instructive to look at the NDCG as well. Here, the baseline model achieves an NDCG of around 0.9 only 5 hours after publication.

In addition to achieving the best model performance in our experiments, the RF model also gives us a natural way to quantify the **prediction uncertainty**. Instead of just calculating the mean of the ensemble predictions, we can calculate percentiles of the predictions to get prediction intervals with 95% coverage. This is shown in two example time series in Fig. 5.

Finally, the trained AR model is **deployed** to do online predictions on real-time data. The data extraction server is deployed on the Google Cloud, which constantly retrieves new tweets and articles and finds the tweet-article matchings. The matchings are saved as a new batch into a document database, deployed on Amazon Web Services. The online news popularity prediction is visualized in an interactive webpage[1], which also provides the open sourced datasets collected by the pipeline.

[1]Link to the webpage: http://cryptodatathon.com/ranknews

6 CONCLUSION

In this paper, we introduce an online data mining system relating cryptocurrency news to the tweets discussing them. This data pipeline paves the way for monitoring cryptocurrency news of public's interest, identifying and predicting popular news, and tracking public opinions towards cryptocurrencies.

Data exploration on the collected paired news articles and tweets characterized top publishers, and top cryptocurrencies discussed on Twitter. We also perform preliminary predictive analytics using machine learning and deep learning models. This work is a first step towards providing a prediction system, that detects articles that are going to become popular shortly after they are published.

Acknowledgments

N.A.-F. and T.G. are grateful for financial support from the EU Horizon 2020 projects: SoBigData under grant agreement No. 654024. J.B., R.H. and D.L are grateful for the support of Professor A. Krause on the Data Science Lab 2018 course at ETH Zurich.

REFERENCES

[1] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Clara Vega. 2007. Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics* 73, 2 (2007), 251–277.
[2] Ioannis Arapakis, B Barla Cambazoglu, and Mounia Lalmas. 2014. On the feasibility of predicting news popularity at cold start. In *International Conference on Social Informatics*.
[3] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The pulse of news in social media: Forecasting popularity. *ICWSM* 12 (2012), 26–33.
[4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[5] Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula, and Lakshminarayanan Subramanian. 2016. Predicting socio-economic indicators using news events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1455–1464.
[6] Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2012. Customers as Advisors: The Role of Social Media in Financial Markets. *SSRN Electronic Journal* (2012).
[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv e-prints* abs/1412.3555 (2014).
[8] Hongjian Dou, Wayne Xin Zhao, Yuanpei Zhao, Daxiang Dong, Ji-Rong Wen, and Edward Y Chang. 2018. Predicting the Popularity of Online Content with Knowledge-enhanced Neural Networks. In *ACM KDD*.
[9] Tian Guo, Albert Bifet, and Nino Antulov-Fantulin. 2018. Bitcoin Volatility Forecasting with a Glimpse into Buy and Sell Orders. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE.
[10] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 261–269.
[11] Ryota Kobayashi and Renaud Lambiotte. 2016. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*.
[12] Nuno Moniz, Luís Torgo, and Fátima Rodrigues. 2014. Resampling approaches to improve news importance prediction. In *International Symposium on Intelligent Data Analysis*.
[13] Matija Piškorec, Nino Antulov-Fantulin, Petra Kralj Novak, Igor Mozetič, Miha Grčar, Irena Vodenska, and Tomislav Šmuc. 2014. Cohesiveness in Financial News and its Relation to Market Volatility. *Scientific Reports* 4, 1 (2014).
[14] Bichen Shi, Georgiana Ifrim, and Neil Hurley. 2014. Insight4news: Connecting news to relevant social conversations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 473–476.
[15] Robert H. Shumway and David S. Stoffer. 2017. *Time Series Analysis and Its Applications*. Springer International Publishing.
[16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.
[17] Manos Tsagkias, Maarten De Rijke, and Wouter Weerkamp. 2011. Linking online news and social media. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 565–574.
[18] Wenbin Zhang and Steven Skiena. 2010. Trading strategies to exploit blog and news sentiment. In *In Fourth Int. Conf. on Weblogs and Social Media (ICWSM)*.

User Donations in a User Generated Video System

Adele Lu Jia
China Agricultural University
Beijing, China
adele.lu.jia@gmail.com

Xiaoxue Shen
China Agricultural University
Beijing, China
xiaoxueshen@cau.edu.cn

Siqi Shen*
National University of Defense Technology
Changsha, China
shensiqi@nudt.edu.cn

Yongquan Fu
National University of Defense Technology
Changsha, China
yongquanfu@nudt.edu.cn

Liwen Peng
National University of Defense Technology
Changsha, China
pengliwen13@nudt.edu.cn

ABSTRACT

User generated video systems like YouTube and Twitch.tv have been a major internet phenomenon. They have attracted a vast user base with their many and varied contents provided by their users, and a series of social features tailored for online viewing. In hoping for building a more lively community and encouraging the content creators to share more, recently many such systems have introduced crowdsourcing mechanisms wherein creators get tangible rewards through user donations. User donation is a very special form of user relationships. It influences user engagement in the community, and has a great impact on the success of these systems. However, user donations and donation relationships remain trade secrets for most enterprises and to date are still unexplored. It is not clear at what scale are the donations or how users donate in these systems. In this work, we attempt to fill this gap. We obtain and provide a publicly available dataset on user donations in BiliBili, a popular user generated video system in China with 76.4 million average monthly active users. Based on detailed information on over 5 million videos, over 700 thousand content creators, and over 1.5 million user donations, we quantitatively reveal the characteristics of user donations, we examine their correlations with the upload behavior and content popularity of the creators, and we adopt machine-learned classifiers to accurately predict the creators who will receive donations and who will donate in the future.

CCS CONCEPTS

• **Information systems** → *Multimedia information systems*; *Collaborative and social computing systems and tools.*

KEYWORDS

User donation, User Generated Content (UGC), user generated video systems

*Shen and Jia serve as the corresponding authors of this article.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316702

ACM Reference Format:
Adele Lu Jia, Xiaoxue Shen, Siqi Shen, Yongquan Fu, and Liwen Peng. 2019. User Donations in a User Generated Video System. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316702

1 INTRODUCTION

User generated video systems nowadays entertain over a billion users and form a billion-dollar global industry. Their success highly depends on the contents provided by their users, and a variety of social features deployed for their users to interact. To keep their proliferation, recently many such systems, for example, YouTube and Twitch.tv, have introduced crowdsourcing mechanisms and many content creators have actually attracted other users to donate [4, 5]. Being a very special form of user relationships that requires real effort to be established and maintained, user donation directly reflects and influences user engagement in the community, and therefore has a great impact on the success of user generated video systems.

However, user donation is considered trade secrets in most enterprises and to date remains an unexplored area. In this article we conduct, to the best of our knowledge, the first in-depth analysis of user donations in user generated video systems. The main motivation behind is to fill the gap between the profound and the promising role that user donation plays in real user-generated video systems and the very limited understanding of it in academia.

To this end, we have chosen BiliBili [3] as our research platform. BiliBili is a popular user generated video systems for the young generations in China. Similar to YouTube, It provides both video and social services. Since it was first launched in July 2009, BiliBili has attracted a large number of users (with 76.4 million average monthly active users for 2018) and issued their IPO for a total offer amount of 483 million dollars in March 2018 [14]. On January 15th, 2016 (much earlier than YouTube), BiliBili introduced a crowdsourcing project. Any interested creators could sign up and receive donations from other users. Unlike YouTube and Twitch.tv wherein user donation remains a trade secret and the donation statistics are not publicly available, for each creator, BiliBili displays the number of donations they have received, along with a list of the identities of the top 30 donors. By doing so, BiliBili creates a sense of friendly competition and encourages its users to donate.

Our analysis of user donations mainly consists of the following three parts:

Measuring user donations. To the best of our knowledge, we provide the first large-scale measurement and publicly available dataset on user donations in user generated video systems (Section 2). Our dataset covers 5,992,355 videos and 734,202 creators. The information we obtained includes not only basic video characteristics like the duration and the popularity, but also user activities and interactions like how users follow and donate to each other, who uploads which video and how these videos perform. For academic purposes, our dataset is publicly available upon request.

Characterizing user donations. We first quantitatively reveal the scale of user donations in BiliBili (Section 3.1). We observe over 1.5 million donations in total and over 20 thousand donations in the month of our crawling alone. While most creators in BiliBili do not receive any donations or only receive a few donations, we find that over 30% of the total donations are destined to 526 creators who have signed and declared to share content exclusively in BiliBili, indicating that exclusive user generated contents are deeply appreciated.

We then dissect the donation composition and reveal the origins of the donations (Section 3.2 and 3.3). We find that the majority (79.37%) of the donations are from viewers who have not shared any contents. Surprisingly, a considerable amount (3.99%) of donations are self-donations, which are probably used to encourage others to donate. From the donor's perspective, we find that the majority of donors only donate once whereas a few donors have returned and donated multiple times within one month.

Finally, we analyze the correlations between user donation and the upload activity and content popularity of the creators (Section 3.4). Although without qualitative analyses like surveys and interviews we cannot argue the causations, we do find that in general creators that have joined the crowdsourcing project (with and without actual donations) are more active and have shared more contents, and that for each video they share, creators with actual donations collected more views than other creators. When we compare the upload activity of the creators before and after BiliBili introduced the crowdsourcing project, we find that those who upload more frequently have received more donations.

Predicting user donations. Applying our findings, we build machine-learned classifiers to predict, without using any information on past donations, the creators *who will receive donations* and *who will donate* in the future (Section 4). On a balanced dataset our predictions achieve an accuracy of 83% and 79% for the two tasks, respectively. Our models provide insights for communities that are considering deploying crowdsourcing mechanisms and can be used to identify in advance the creators that will receive donations and the potential donors.

2 THE BILIBILI DONATION DATASET

In this section, we give a brief introduction of the ecosystem of BiliBili, and we introduce our measurement methodology and the dataset used throughout this article.

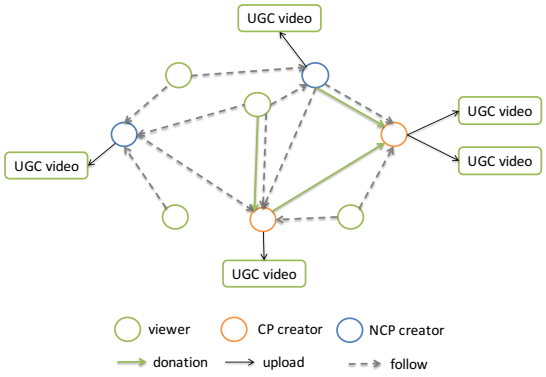


Figure 1: The BiliBili ecosystem. Three types of users include creators that have joined the crowdsourcing project (named *CP creators*), creators that did not join the crowdsourcing project (named *NCP creators*), and users who have not uploaded any contents (named *viewers*). Three types of links represent the *donation* relationships, the *follow* relationships, and the *upload* relationships, respectively.

2.1 The BiliBili ecosystem

BiliBili provides both video and social services. As in traditional user generated video systems like YouTube, users in BiliBili can share and view videos, vote and leave comments to videos, and subscribe to other users. In addition, BiliBili provides several enhanced social features, such as social network incorporation, and chat replay (named *danmu* in BiliBili) wherein the chat from the past show up right next to and on top of the video for the current viewer.

The majority of BiliBili creators are regular users with no affiliations, while a small number of the creators are *signed creators*, i.e., creators who sign up with BiliBili and declare to share contents exclusively in BiliBili, and *branding accounts*, i.e., organisations that share contents for promotion. On January 15th, 2016, BiliBili introduced a *Crowdsourcing Project* (CP) wherein any interested creators could sign up and receive donations from other users (possibly). By default, the number of donations the creators received, in total and for the current month, are highlighted in their home-pages, along with a list of the top donors.

Figure 1 shows an overview of the BiliBili ecosystem.

2.2 The BiliBili donation dataset

BiliBili identifies each of its video and each of its users with a unique number in the increasing order. Each identifier corresponds to a webpage with detailed video or user information that is publicly accessible and can be obtained with web crawlers. To give a sufficient observing period, we focus on creators that have joined BiliBili before May 2017. In total, we have obtained 136,375 and 597,827 creators that have and have not joined the crowdsourcing project, which we name *CP creators* and *NCP creators*, respectively.

For each of the CP creators, we obtain information on (i) the total number of donations he received, (ii) the number of donations and the list of the top 30 donors in the month of our crawling (April 2018), and in addition (iii) the gender, the register time, the list of users that he follows, and the creator type (regular, signed, or

Table 1: Basic statistics of the BiliBili donation dataset.

#CP creators total/signed/brand	136,375/526/325
#CP creators male/female	57,837/23,067
#NCP creators	597,827
aggregate #donations (total)	1,561,655
aggregate #donations (recent month)	51,960
#donation relationships (recent month)	25,265
#following relationships	10,543,151
#videos	5,992,355
aggregate video length	239 years
aggregate viewing time	5.37 million years
aggregate #views	50 billion
mean #views	8.468

brand). Overall, the CP creators have aggregated over 1.5 million donations in total and over 50 thousand donations in the month of our crawling alone. A very small fraction of them are signed creators (526) and brand accounts (325). For creators who choose to reveal their gender (which is unknown by default), we find 57,837 male creators and 23,067 female creators, respectively.

We obtain the upload activity and the content popularity of both CP and NCP creators through crawling the video pages. Our crawling was carried out in April and May 2018. To give a sufficient time for each video to collect its popularity, we have only considered videos that were uploaded at least one month ago, i.e., until March, 2018, since according to many studies on user generated video systems [6, 7, 10], video popularity rarely changes one month after the upload. In total, we have obtained detailed information on 5,992,355 videos including, for each video, (i) the uploader, (ii) the duration, (iii) the category (e.g., gaming, life, music, etc.), and (iv) the popularity attributes including the number of views, the number of favorites, the number of danmus, and the number of comments it collected.

The basic statistics of our datasets are depicted in Table 1. To give context regarding the types of the content, we briefly investigate the video category and the categories of the CP creators. Here, we choose the category of the videos that a creator has uploaded most as the category of the creator. Overall, 34.75% CP creators have uploaded videos solely from one category and the dominant category of 63.54% CP creators exceeds half of their uploads. Figure 2 shows the fraction of CP creators that fall into each category, as well as the fraction of the aggregate number of views and donations collected within each category. The gaming category, including game replays and derivatives, attracts the largest fraction of creators, views, and donations.

3 UNDERSTANDING USER DONATIONS IN BILIBILI

In this section, we reveal the basic characteristics of user donations in BiliBili.

3.1 The number of donations received

We begin by revealing the number of donations received by the 136,375 CP creators. Overall, the CP creators have aggregated 1,558,932 donations in total and 50,667 donations in the month

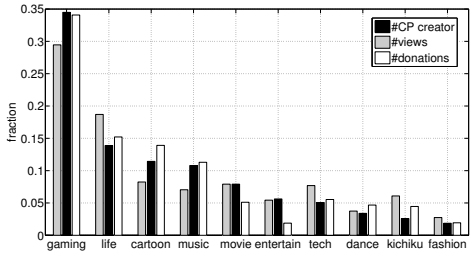


Figure 2: Fraction of CP creators, views, and donations for each category. BiliBili provides 17 categories and here we show the top 10 categories in terms of the number of CP creators. In total, it covers 97% of all the CP creators.

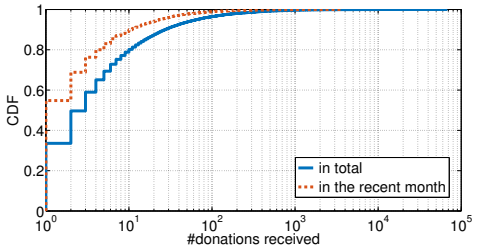


Figure 3: CDFs of the number of donations received in total and in the recent month. Note that the horizontal axis is in log scale.

Table 2: Statistics on the number of donations received in total and in the recent month for different groups of creators. *User base* shows the number of creators that have received at least one donation in each creator group. *Fraction* shows the fraction of total donations aggregated by each creator group.

total	**user base**	**fraction**	Q_1	**mean**	Q_3	**max**
regular	51,164	60.91%	1	17	7	13,300
signed	517	30.17%	65	910	807	39,900
brand	262	8.78%	6	307	122	17,800
month	**user base**	**fraction**	Q_1	**mean**	Q_3	**max**
regular	5,648	57.93%	1	5	3	1,020
signed	317	34.25%	2	55	28	3,740
brand	85	6.52%	2	39	20	1,160

of our crawling alone. This result quantifies that the crowdsourcing project is running actively and is well accepted by BiliBili users.

At the individual level, however, we find that only 51,989 (38.12%) CP creators were able to actually receive any donations, among which 6,057 (11.65%) managed to receive donations in the recent month. On average, they collected 30 donations in total and 8 donations in the recent month, respectively. As further depicted in Figure 3, we observe disparities for both donation measures. Particularly, 201 creators and 66 creators have attracted over 1,000 donations in total and over 100 donations in the recent month respectively, whereas 17,457 creators have received donations only once.

Creator type. Among the 136,375 CP creators, we find 526 *signed creators*, i.e., creators who have signed up and declared to

Table 3: Donation and donator composition

donation composition	from CP creators	from NCP creators	from viewers	self-donation
count	726	3,447	20,053	1,009
fraction	2.87%	13.64%	79.37%	3.99%
donor composition	**CP creators**	**NCP creators**	**viewers**	
user base	136,375	597,827	NA	
donor count	1,669	3,312	19,293	
donor fraction	1.22%	0.28%	NA	

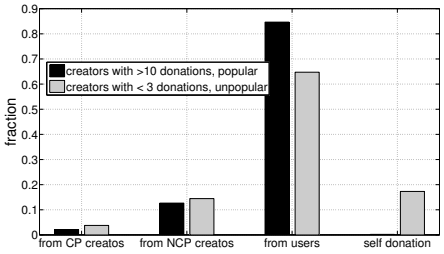


Figure 4: Donation composition: the fraction of donations received in the recent month from different groups of users for popular creators (with more than 10 donations) and unpopular creators (with fewer than 3 donations).

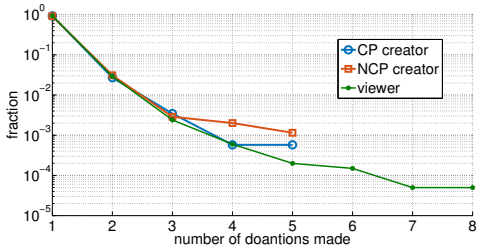


Figure 5: Donor composition: the fraction of donors by the number of donations made in the recent month. Note that the vertical axis is plotted in the log scale. In total, 96.55% donors have donated once and 3.45% donors return for a second donation within one month. Particularly, 5 donors have donated to more than five creators in the recent month and they are the viewers.

share contents exclusively in BiliBili, and 325 *brand accounts*, i.e., organizations that share contents for promotion. Intuitively, they are expected to be more professional than regular creators with no affiliations. As shown in Table 2, clearly signed creators and the brand accounts receive much more donations than regular creators. Surprisingly, while both being relatively more professional, signed creators attract roughly three times more donations than the brand accounts. And although being a minority, they have accumulated over 30% of the total donations, suggesting *deep appreciations of BiliBili viewers towards exclusive contents.*

3.2 Where do the donations come from?

The previous section has revealed that donations are conducted very actively in BiliBili. Here, we examine the donation relationships and reveal the origins of the donations. BiliBili displays for each creator the top 30 donors in the recent month. In total, we find that 6,057 creators have received donations in the month of our crawling, among which 5,816 (96.01%) have received no more than 30 donations, i.e., we have captured all the recent donation relationships for these creators and the top 30 donors (representing the most supportive ones) for the rest.

Donation composition. Depending on the origins, we decompose donations into four categories, namely donations received *from CP creators*, *from NCP creators*, *from viewers*, and *self-donations*, respectively. Reciprocal donations are rare in BiliBili. Among the 6,057 creators, we only find 8 user pairs that have donated mutually. Overall, the four donation categories each represents 2.87%, 13.76%, 79.37%, and 3.99% of the total donations, respectively. Not surprisingly, most of the donations are made by the viewers, possibly due to the large user base of viewers compared to creators. Interestingly, we find a non-neglectable fraction of donations are self-donations. We conjecture that users are encouraging others to donate by "dropping the first coin" like the street artists usually do.

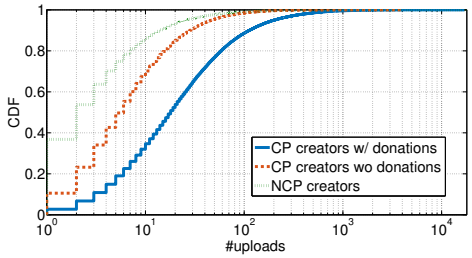
To take a closer look, we differentiate the following two groups of creators, i.e., creators with fewer than three donations (*unpopular*, in terms of donations) and creators with more than 10 donations (*popular*). Figure 4 shows the donation composition of these two types of creators, respectively. The major differences in the donations they received are that (i) unpopular creators have performed a lot more self donations; and (ii) popular creators achieve a much larger fraction of donations from viewers. It seems that "dropping the first coin" is not as effective as the unpopular creators thought.

Donor composition. The above analysis reveals the composition of donations received by each individual creator. Here, we take one step back and examine at the community level the composition of the donors. Our dataset captures in total 136,375 CP creators and 597,827 NCP creators, among which we find that 1,669 CP creators and 3,312 NCP creators have performed donations, representing 1.22% and 0.28% of the corresponding user base, respectively. Clearly, CP creators are more supportive and engaging in the crowdsourcing project.

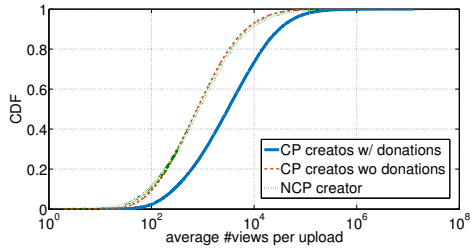
Table 3 summarizes the basic statistics for the above analysis.

3.3 Donor returns?

Finally, we focus on the donors and reveal whether donors return to make another donation. Figure 5 shows the fraction of donors that donate once to 8 times (the maximum) in the month of our crawling. Overall, 96.55% donors have donated once and only 3.45% donors return for a second donation within one month. Separating CP creators, NCP creators, and viewers, we find that a slightly higher



(a) number of uploads



(b) average number of views per uploads

Figure 6: Creator activity and popularity. Here we have divided the creators into three groups, i.e., CP creators who have received at least one donation, CP creators who did not receive any donation, and NCP creators.

fraction (3.90%) of NCP creators have returned, nevertheless, only viewers have returned for more than 5 times. The low donor return rate is possibly due to the small observing period in our experiment. For our future work, we plan to follow the creators for a longer period for verification.

3.4 Correlations with user engagement

In previous sections, we have revealed the characteristics of user donations. Here, we further analyze their correlations with user engagement. We focus on two fundamental user engagements in BiliBili, namely the creator's willingness to share new contents and the viewer's participation reflected by the content popularity (e.g., number of views). Particularly, we seek to quantitatively answer the following two questions:

1. *How are user donations correlated with the upload activity and the content popularity of the creators?*

2. *Do creators change their upload behavior after BiliBili introduced the crowdsourcing project?*

Q1: To answer the first question, we divide the creators into three groups, namely CP creators with donations, CP creators without donations, and NCP creators. We begin by examining the upload activity for creators in each group. As shown in Figure 6(a), CP creators with donations are the most active in uploading. On average they have uploaded 56 videos whereas for NCP creators it is 7.19. Interestingly, CP creators with no donations, i.e., those who have joined the crowdsourcing project but have not received any donations, are more active than NCP creators, i.e., those who did not join the crowdsourcing project.

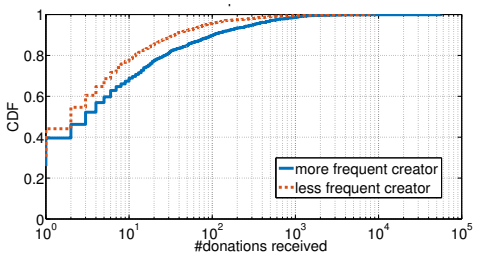


Figure 7: CDF of the number of donations received by creators that upload more and less frequently since the crowdsourcing project was introduced.

The above results quantify the correlations between joining the crowdsourcing project, receiving donations, and the upload activity of the creators. The causation, on the other hand, is difficult to argue quantitatively. It is possible that joining the project alone stimulate creators' willingness to share new contents. It is also possible that active creators are in the first place more willing to join the crowdsourcing project. Further analysis will depend on qualitative analysis such as surveys and interviews, which we leave for our future work.

Figure 6(b) further shows the CDFs of the content popularity, in terms of per-upload number of views, collected by the three creator groups. We find that CP creators with donations are in general more popular while CP creators with no donations and NCP creators achieve similar popularity. This result quantifies the correlation between viewers' participation and receiving donations, but different from the above analysis, simply joining the project does not make a difference.

Q2: To answer the second question, we compare the upload frequency of CP creators before and after BiliBili introduced the crowdsourcing project. We focus on the 4,008 CP creators who have uploaded at least 10 videos for both periods, and find that 2,785 creators have uploaded less frequently (with a longer average interval between uploads) since the crowdsourcing project was introduced, whereas 1,223 creators have uploaded more frequently. Naming them *less frequent creators* and *more frequent creators* respectively, in Figure 7 we show the distributions of the number of donations that they receive. we find that more frequent creators in general tend to get more donations.

Intuitively, there exist many possible reasons for the change in upload activities. Internally, creators could simply get bored or more engaged to the community. Externally, they could be stimulated by the feedback of the viewers (views, donations, etc.). It is possible that their high activity level exposes them to more potential donors, or that receiving donations encourage them to be more active. Similar to the **Q1**, our dataset can only quantify the correlations between them. Arguing the causation requires qualitative analysis such as surveys and interviews which we leave for our future work.

4 PREDICTING USER DONATIONS

Having gained valuable insights from the previous section, in this section we predict user donations using standard machine learning techniques. We consider two prediction tasks, namely, based on the

upload and the social activity of the creators, to predict (i) *which creator will receive donations* and (ii) *which creator will donate* in the following month. Both tasks shed light on community maintenance. Particularly, the former one reveals by whom and to what extent users are attracted to the community. The latter one helps identifying potential donors whose donation will be a tangible reward for the creators.

Experimental setup. To this end, we first took a snapshot of the BiliBili community and obtained the upload activities of the creators and their content popularity until the end of March 2018. Then we focused on the 136,375 CP creators captured in our dataset and observed them for one month. During this observation period, we recorded 6,220 creators that have received donations and 4,981 creators that have donated to them. We label them as positive examples for the above-mentioned prediction tasks, respectively.

Based on previous analysis, we extract three groups of features including creator attributes (C), upload activities (U), and the follow graph properties (G). Features for the former two groups have been extensively studied in Section 3. In addition, we include features extracted from the follow graph, which is constructed by the following relationships, to test whether structural properties provide additional information on the prediction tasks. All features are summarized in Table 4.

We experimented with a variety of classification algorithms—logistic regression, SVM, and random forests—and found their performance similar. Hence all results reported here were obtained using SVM. We use balanced training and test sets containing equal numbers of positive and negative examples. For each experiment, we run 5-fold cross-validation and report the area under the receiver operating characteristic (ROC) curve (AUC).

Results. The prediction results are shown in Table 5. We have a number of interesting findings as follows.

First, consistent with the analysis in Section 3, the upload activities and content popularity of the creators (U) are very informative for predicting whether they will receive donations in the future (Task 1), however, they are not as effective for predicting who will donate (Task 2). Creator attributes, on the other hand, provide more information for Task 2. Secondly, for Task 1, the follow graph features are comparable and sometimes even more useful, showing a strong correlation between the social status of the creators and user donations. Thirdly, combining all features together achieves the best performance, and predicting who will receive donations is relatively easier than predicting who will donate: the former one achieves an AUC of 0.83 whereas for the latter task it is 0.79.

More specifically, the top three most informative features for predicting who will received donations (Task 1) are the number of followers of the creator, the number of views collected, and the number of videos uploaded previously by the creator, among which the number of views are negatively related while the rest are both positively related to Task 1. While it seems natural that creators with a large number of followers and/or have uploaded many videos are more likely to received donations in the future, it is surprising to find out that the number of views collected by the creators in fact has a negative influence on whether they will receive donations. We conjecture that viewing a video does not directly reflect user appreciation. Users may simply be exploring, or they may dislike the video and will quickly turn it off. For predicting which creators will donate (Task 2), the top three most informative features are the user type (CP or NCP creators), the number of views collected, and the number of users that creators follow. All the three features are positively related to Task 2. Different from the result for Task 1, here, the number of views creators received has a positive influence on whether they will donate to other creators. We conjecture that many creators consider receiving a large number of views as a reflection of being well accepted by the community, and they might be willing to "return the favor" in some forms including donation.

Table 4: Prediction features. We consider three groups of features and their combinations including creator attributes (C), upload activities (U), and follow graph properties (G)

group	feature description
C	gender, type, register time
U	number of uploads number of views, shares, comments, and danmus collected
G	number of followers and followees clustering coefficient, and PageRank score

Table 5: Performance evaluation (based on AUC) of predicting which creators will receive donations (Task 1) and which creators will donate (Task 2).

feature	Task 1	Task 2
C	0.5967	0.7458
U	0.8089	0.6581
G	0.7915	0.7193
$C + U$	0.8221	0.7552
$C + U + G$	0.8304	0.7852

Discussion. In the above prediction tasks, we did not use any information on the past donations, which will clearly improve the performance of our models, for the reason that we seek to infer user donations solely from the upload and the social activities of the creators. In this way, we provide insights for communities that are considering deploying crowdsourcing projects and our models can be used to identify in advance the creators that will receive donations and the potential donors. It should also be noted that for our prediction tasks we have omitted viewers, who have not shared any contents and therefore we have very limited information on them. Clearly, there exist many possible ways to improve our prediction models and also to include viewers, for example, through incorporating finer-grained network features learned from network representation learning models that are shown to perform well in node classification and link prediction tasks [13]. The key to this method is to propose new models that can cope with the heterogeneous multi-view relationships in BiliBili. We leave this as our future work.

5 RELATED WORK

We summarize related work within each research topic our work covers as follows:

User generated video systems. User generated video systems like YouTube and Twitch.tv have been extensively studied before. Cha *et al.* presented a comprehensive analysis of the popularity distribution and the time evolution of UGC video requests and their

implications [6]. Ding *et al.* analyzed in-depth the behaviors of YouTube uploaders [9]. Gill *et al.* investigated YouTube from the perspective of YouTube traffic [11]. They examined YouTube usage patterns, file properties, and transfer characteristics. Kaytoue *et al.* provided preliminary characterizations on Twitch. They analyzed the dynamics of game spectators and proposed models for predicting video popularity [18]. Wattenhofer *et al.* analyzed the correlations between the popularity of YouTube videos and the properties of various social graphs created among the users [24]. In our previous works, we compared Twitch with other systems and investigated their repositories and user activities [16], and we analysed the content popularity in BiliBili and leveraged social features for user activity prediction [15, 17].

Different from the above studies, our analysis of video systems focuses on a very special form of user relationships, i.e., user donation, which to the best of our knowledge is still unexplored.

Crowdfunding systems. On the other hand, user donation in crowdfunding platforms, wherein entrepreneurs solicit funding in order to bring their business plans, have been analyzed before, ranging from predicting the success of crowdfunding campaigns [8, 12, 20, 25] to investor and project recommendation [2, 21, 22], group recommendation [23], donor retention [1] and competition modelling [19]. They mainly rely on probabilistic generative models or manual feature engineering (based on profile and social features) to build machine-learned classifiers to predict the success and the potential investors of the projects.

The above studies mainly focus on crowdfunding platforms in the context of raising money for commercial projects, wherein donors are reimbursed by receiving interests or by pre-ordering the products. Being a user generated video system, BiliBili provides a completely different context, and moreover, donors in BiliBili do not expect any real return except for some videos and a friendly community. Thus, the donor dynamics and the donation relationships are expected to be different are worth investigating.

6 CONCLUSION AND FUTURE WORK

In this work, we conducted an analysis and presented the first publicly available dataset on user donations in user generated video systems. Based on statistics on over 130 thousand creators who have joined the crowdsourcing project in Bilibili, we investigated the dynamics of user donations and we applied our findings to accurately predict future donations.

We have a number of interesting findings. First, among the 130 thousand creators, a few hundred signed creators, those who declared to share exclusively in BiliBili, have accumulated over 30% of the total donations, showing deep appreciations of BiliBili users towards exclusive contents. Secondly, the majority (around 80%) of the donations are from viewers and a considerable amount (4%) of donations are self-donations. Thirdly, we observe that joining the crowdsourcing project (even with no donations) is correlated with the upload activity but not with the content popularity of the creators. Finally, using simple features extracted from the upload activity, the popularity and the social status of the creators, we can predict with high accuracies the future donations.

As stated throughout this article, our work can be improved in a number of ways. The first would be qualitative analyses for arguing the causation on user donations. So far we have quantified (strong) correlations between user donations and the upload activity and the content popularity of the creators. It would be beneficial for the communities to understand the motivations behind, so that specific incentive policies could be designed to encourage users to donate, and creators in return will be motivated to share more. Secondly, we observe that a large fraction of donations come from viewers, on whom however we have very limited information and therefore cannot make predictions on their future donations. Relationships that they involve in, for example, social and donation relationships, would be valuable supplements and could be explored based on network representation learning models. The key to this method would be proposing suitable models that can incorporate the heterogeneous multi-view relationships in BiliBili.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation for Young Scholars of China (NSFYSC) No. 61502500 and No. 61602500.

REFERENCES

[1] T. Althoff and J. Leskovec. 2015. Donor Retention in Online Crowdfunding Communities: A Case Study of DonorsChoose.org. In *Proceeding of the 24th International World Wide Web Conference (WWW'15)*.
[2] J. An, D. Quercia, and J. Crowcroft. 2014. Recommending investors for crowdfunding projects. In *Proceeding of the 23th International World Wide Web Conference (WWW'14)*.
[3] BiliBili. 2017. https://www.bilibili.com.
[4] Twitch Official Blog. 2017. https://blog.twitch.tv/twitch-affiliate-program-launches-today-first-invites-going-out-9c020994a042.
[5] YouTube Official Blog. 2017. https://youtube.googleblog.com/2017/01/can-we-chat-hello-super-chat.html.
[6] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. 2007. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Internet measurement conference (IMC'07)*.
[7] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. 2009. Analyzing the Video Popularity Characteristics of Large-scale User Generated Content Systems. *IEEE/ACM Transactions on Networking* 17, 5 (2009), 1357–1370.
[8] J. Chung and K. Lee. 2015. A long-term study of a crowdfunding platform: Predicting project success and fundraising amount. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*.
[9] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K.W. Ross, and A. Ghose. 2011. Broadcast Yourself: Understanding YouTube Uploaders. In *Proceedings of the 5th Internet Measurement Conference (IMC'11)*.
[10] F. Figueiredo, F. Benevenuto, and J.M. Almeida. 2011. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the 4th ACM international conference on Web search and data mining (WSDM'11)*.
[11] P. Gill, M. Arlitt, Z. Li, and A. Manhanti. 2007. YouTube Traffic Characterization: A View from the Edge. In *Proceedings of the first Internet Measurement Conference (IMC'07)*.
[12] M. Greenberg, K. Hariharan, E. Gerber, and B. Pardo. 2013. Crowdfunding Support Tools: Predicting Success & Failure. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'13)*.
[13] W.L. Hamilton, R. Ying, and J. Leskovec. 2018. Representation Learning on Graphs: Methods and Applications. In *arXiv:1709.05584*.
[14] BiliBili IPO. 2018. https://www.nasdaq.com/markets/ipos/company/bilibili-inc-1047286-86197.
[15] Adele Lu Jia, Siqi Shen, Shengling Chen, Dongsheng Li, and Alexandru Iosup. 2017. An Analysis on a YouTube-like UGC site with Enhanced Social Features. In *Proceedings of the 26th International World Wide Web Conference (WWW'17 Companion)*.
[16] Adele Lu Jia, Siqi Shen, Dick Epema, and Alexandru Iosup. 2016. When Game Becomes Life: The Creators and Spectators of Online Game Replays and Live Streaming. *ACM Trans. on Multimedia Computing, Communications, and Applications* 12, 4 (August 2016).
[17] Adele Lu Jia, Siqi Shen, Dongsheng Li, and Shengling Chen. 2018. Predicting the implicit and the explicit video popularity in a User Generated Content site with enhanced social features. *Computer Networks* 140 (2018), 112–125.
[18] M. Kaytoue, A. Silva, and C. Raissi. 2012. Watch me Playing, I am a Professional: a First Study on Video Game Live Streaming. In *Proceedings of the 12th International*

World Wide Web Conference (WWW'12 Companion).

[19] Y. Lin, P. Yin, and W.C. Lee. 2018. Modeling Dynamic Competition on Crowdfunding Markets. In *Proceeding of the 27th International World Wide Web Conference (WWW'18).*

[20] C.T. Lu, S. Xie, X. Kong, and P. S. Yu. 2014. Inferring the impacts of social media on crowdfunding. In *Proceedings of the 7th ACM international conference on Web Search and Data Mining (WSDM'14).*

[21] V. Rakesh, J. Choo, and C.K. Reddy. 2015. Project Recommendation Using Heterogeneous Traits in Crowdfunding. In *Proceeding of the 24th International World Wide Web Conference (WWW'15).*

[22] V. Rakesh, J. Choo, and C. K. Reddy. 2015. Project recommendation using heterogeneous traits in crowdfunding. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (CWSM'15).*

[23] V. Rakesh, W.-C. Lee, and C. K. Reddy. 2016. Probabilistic group recommendation model for crowdfunding domains. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16).*

[24] M. Wattenhofer, R. Wattenhofer, and Z. Zhu. 2012. The YouTube Social Network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12).*

[25] A. Xu, X. Yang, H. Rao, W.-T. Fu, S.-W. Huang, and B. P. Bailey. 2014. Show me the money!: an analysis of project updates during crowdfunding campaigns. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'14).*

CrowdGuard: Characterization and Early Detection of Collective Content Polluters in Online Social Networks

Ke Li
Northwestern Polytechnical University
Xi'an, Shaanxi, China,
sfk100200@qq.com

Bin Guo
Northwestern Polytechnical University
Xi'an, Shaanxi, China,
guob@nwpu.edu.cn

Qiuyun Zhang
Northwestern Polytechnical University
Xi'an, Shaanxi, China,
1975642935@qq.com

Jianping Yuan
Northwestern Polytechnical University
Xi'an, Shaanxi, China,
jyuan@nwpu.edu.cn

Zhiwen Yu
Northwestern Polytechnical University
Xi'an, Shaanxi, China,
zhiwenyu@nwpu.edu.cn

ABSTRACT

Recently, content polluters post malicious information in Online Social Networks (OSNs), which is a more and more serious problem that poses a serious threat to the privacy information, account security, user experience, etc. They continuously simulate the behaviors of legitimate accounts in various ways, and evade detection systems against them. In this paper, we focus on one kind of content polluter, namely collective content polluter (hereinafter referred to as CCP). Existing works either focus on individual polluters or require long periods of data records for detection, making their detection methods less robust and lagging behind. It is thus necessary to analyze the characteristics of collective content polluters and study the methods for early detection. This paper proposes a CCP early detection method called CrowdGuard. It analyzes the crowd behaviors of collective content polluters and legitimate accounts, extracts distinctive features, and leverages the Gaussian Mixture Model (GMM) method to cluster the two groups of accounts (legitimate users and polluters) to achieve early detection. Using the public dataset including thousands of collective content polluters on Twitter about a political election, we design an experimental scenario simulating early detection and evaluate the performance of CrowdGuard. The results show that CrowdGuard outperforms existing methods and is adequate for early detection.

CCS CONCEPTS

• Information systems ~ Clustering • Information systems ~ Social networks • Security and privacy ~ Social network security and privacy • Computing methodologies ~ Mixture modeling

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316452

KEYWORDS

Social Media, Collective Content Polluters, Early Detection, Crowd Computing, Gaussian Mixture Model

ACM Reference format:

Ke Li, Bin Guo, Qiuyun Zhang, Jianping Yuan and Zhiwen Yu. 2019. CrowdGuard: Characterization and Early Detection of Collective Content Polluters in Online Social Networks. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316452

1 Introduction

While online social networks (OSNs), such as Twitter, Facebook and Linked-In, offer a variety of conveniences and meet people's needs, it has become a profitable platform for attackers because of its huge user base [1]. By creating a large number of fake accounts or hijacking legitimate accounts, attackers post malicious information such as advertisements, pornography and phishing [2]. These fake accounts [3] and compromised accounts [4, 5], especially those that publish malicious information, are called *content polluters*. These malicious behaviors pose a serious threat to legitimate users' privacy information, account security and user experience [6]. Thus, it is necessary to characterize these content polluters in detail and propose automatic detection methods.

Content polluters have different forms, such as social bots [7], Sybil accounts [8], spammers [9], compromised accounts [4, 10], spam campaign [11], and so on. Among them, social bots which are controlled by software are widely adopted by malicious users, since they can be automatically created and generate and disseminate many spam messages easily. With so many forms, these content polluters have different behavior patterns and features, which poses a challenge to research techniques for detecting them.

Facing a variety of detection technology, content polluters must simulate the behaviors of legitimate accounts, so as to

evade detection. Moreover, they always evolve as detection techniques evolve. Recently, a new wave of content polluters is rising [12], called *collective content polluters*. Similar to collective anomalies in anomaly detection field [13, 14], a single CCP mixed in a legit population seems normal, but the occurrence of CCPs together as a collection is anomalous. They perform malicious actions for the same purpose, show similar group behaviors, and are usually event-driven, such as triggered by a malicious mission.

Existing works [15, 16, 17] pay more attention to individual polluter rather than CCP. They generally use the existing data to extract individual features, train a classifier, and then input unknown accounts one by one for detection. However, for CCP, most one-by-one detection methods do not work well, because they cannot capture the group dynamics. Therefore, Cresci et al. [18, 19] study the collective content polluters in a group fashion but their detection method requires long periods of data to distinguish them. Usually, the polluters are detected after the malicious event, and the legitimate users still suffer.

Similar to how law enforcement officers mark and monitor suspects and effectively give early warning before their crimes become apparent [20, 21], in this paper, we propose an *early detection* scenario and a CCP early detection method called *CrowdGuard*. At first, we analyze the group behavior of CCPs and legitimate accounts from multiple perspectives, and get some meaningful findings, which is helpful for early detection of CCP. Then, we extract both individual and group features from different aspects, including user demographics, social network statistics, interaction, content and temporal features. Next, we leverage the Gaussian Mixture Model (GMM) to cluster the two groups (i.e. CCP group and legitimate group, the same below) of accounts to achieve early detection. Finally, using the public dataset about a political election consisting of 991 CCPs and 1083 legitimate accounts on Twitter, we design an experimental scenario simulating early detection and evaluate the performance of our method, CrowdGuard. It achieves more than 98% precision and 94% recall, which indicates that it is adequate for early detection.

We summarize the major contributions of this paper as follows:

(1) We make an in-depth analysis of the differences between collective content polluters and legitimate accounts, and obtain some interesting findings.
(2) We extract 43 individual/group features, and choose GMM to describe these accounts and propose a group detection scheme.
(3) An experimental scenario simulating early detection over a public dataset is designed to evaluate the performance of our method. The results indicate that it is adequate for early detection.

2 Preliminary

We utilize public datasets from [22], part of which are used in our research. The authors focus on some social bots on Twitter during the Mayoral election in Rome in 2014. One of the runners-up hired about a thousand social bots to advertise his policies and help his campaign. After data processing, our existing dataset consists of 991 CCPs and 1083 legitimate accounts. The posting behavior of one CCP is very similar to that of a legitimate account, and they only post a few tweets a day, mainly retweet or copy the tweets of some celebrities. Their profiles are detailed, and it's hard to tell the difference only from their profiles. But the exception is that every time the candidate posts a tweet, they will retweet or copy it within a short period of time, which is just the characteristic of collective content polluters. So, any account that retweets or copies the candidate's tweets will be treated as a suspicious account and its information will be collected. Then 991 social bots will be screened out by manual evaluation, which acts as CCP in our research.

Each account contains user's basic information (friends count, followers count, create time and so on) and its tweets history (each of tweets contains content, URL, mention, hashtag, time stamp and so on) containing up to 3250 recent tweets. That is, one user corresponds to multiple tweets.

Social Fingerprinting [19] is our strong baseline. The authors introduced a bionic technique to model users' online behaviors by so-called "digital DNA" sequence. Each tweet is encoded as a nucleotide base based on its type (simple tweet, reply and retweet), and each account is encoded as a DNA strand (i.e. a string). Intuitively, a group of users who share a longer substring has higher abnormal similarity. To measure their similarity, the authors leverage a generalization of Longest Common Substring (LCS) which can be calculated by a parallel algorithm in [23], and the latent longest DNA substring shared by k accounts in a group is calculated (for each k: $2 \le k \le M$, where M is the group size). Thus, each k value corresponds an *LCS* value, and LCS curve can be plotted (see Figure 1(a)). For legit group, the LCS curve presents an exponential decay trend with the increase of k, and the *LCS* value has dropped to a very low level when k value is small. However, for CCP group, due to the similar behavior pattern, the *LCS* keeps a higher level than that of legit group, and the LCS does not drop sharply until k approaches M. In this way, the LCS curves of the two groups are different. The greater the difference, the easier it is to distinguish the two groups of accounts. Based on this, the detection method is raised. Finally, it can achieve the precision and recall of around 97%.

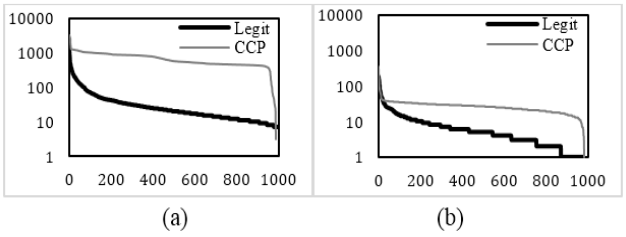


Figure 1: LCS curve in different scenario, where the Y-axis is in log scale. (a) post-event scenario. (b) early scenario.

However, defects still exist. (1) The LCS itself needs to match strings exactly, and the content polluters are judged too strictly.

(2) It is only applicable to long strings (post-event scenario). Figure 1(b) shows the LCS curve in early detection scenario, where most of the DNA chains are short, and the LCS curves of the two groups are less different than those of original scenario. Therefore, abnormal feature fragments are more difficult to be captured, and only a small proportion of content polluters can be detected, resulting in low recall, which will be shown in Section 5. (3) Although group detection is a good idea, the model is too simple and takes too few factors into consideration. Only tweet type or entity information is considered, while content, time and other information are not involved in the model. When applied to the early detection scenario, Social Fingerprinting performs poorly, that is, it is not robust enough.

To address the above defects, we will analyze CCPs in depth, and propose a targeted detection method based on this.

3 Data Analysis and Feature Extraction

3.1 Analysis of Daily Activeness

At first, we define the *daily activeness* of an account m in a group G ($1 \le m \le M$, where M is the group size, and G can be CCP or legit). It is the frequency of all the online activities (simple tweets, replies and retweets) of that account on a given day t, referred to as $A_m(t)$. Then the *daily activeness per capita* is defined as the following formula.

$$APC(t) = [\sum_{m=1}^{M} A_m(t)] / [\sum_{m=1}^{M} I(A_m(t) > 0)]$$

where $I(x)$ is an indicator function of a logical expression x. In fact, $APC(t)$ measures daily activeness of a group. It is worth noting that the $APC(t)$ considers all active users (at least posting one tweet on the day t) rather than all users in G. Figure 2(a) and 2(b) show the daily activeness per capita of CCP group and Legit group between Aug 1, 2014 and Nov 30, 2014, respectively. It can be seen from Figure 2(a) that the daily activeness per capita remains below 10 on the whole, but within 4 days around October 30, it reaches a peak of 106. Therefore, it is speculated that the election event is likely to occur before and after October 30. That is to say, CCPs are less active when there is no event, and their daily activeness can explode in a short time when there is an event. As seen in Figure 2(b), the daily activeness per capita of legitimate accounts fluctuates around 8, with an occasional peak of 9.7. Compared with CCPs, this peak is not exaggerated. It shows that the daily activeness of normal people is random and may be disturbed by some hot events, but the daily activeness of the group tends to be stable over a long period of time.

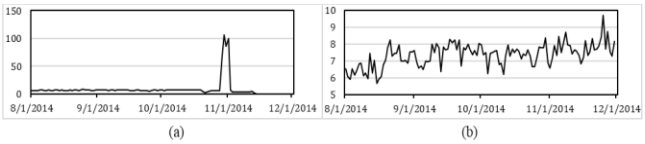


Figure 2: Daily activeness per capita between Aug 1, 2014 and Nov 30, 2014. (a) CCPs. (b) legitimate accounts.

Figure 3(a) and 3(b) give examples of four CCPs' and four legitimate accounts' daily activeness between Aug 1, 2014 and Nov 30, 2014, respectively. From Figure 3(a) we can see that at about day 90 (i.e. 2014/10/30), all of the four CCPs become active suddenly for 4 days, and they are very synchronized, which is just the collective feature. From Figure 3(b) we can see that the daily activeness curve of 4 legitimate accounts is entirely different. As an old saying goes, "No two leaves are identical in the world", which is what legitimate users should look like.

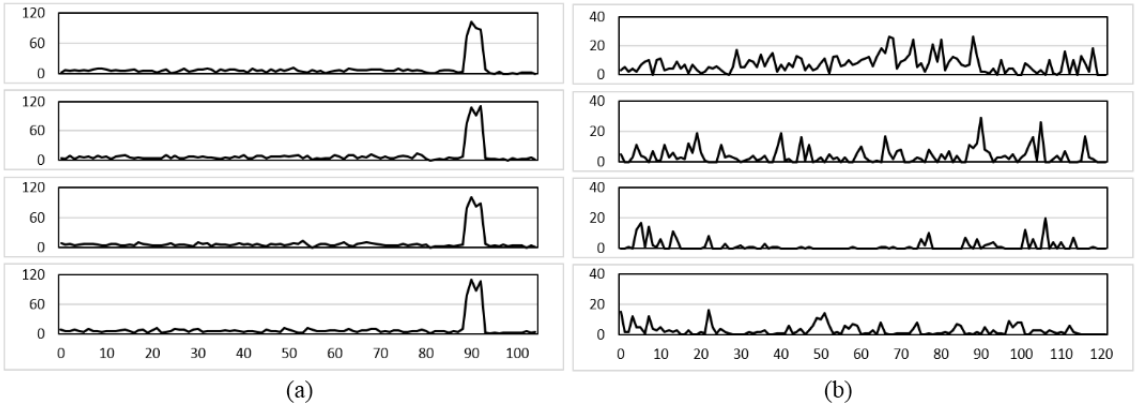


Figure 3: Examples of daily activeness between Aug 1, 2014 and Nov 30, 2014. (a) CCPs. (b) legitimate accounts.

If our application scenario is post-event, then the activeness will be a good indicator for detection. However, our application scenario is early scenario, which should be finished on the day or before the event. So, we also need to analyze the data from other aspects to find out a better indicator.

3.2 Analysis of Online Behaviors

3.2.1 Group Interaction Analysis. It mainly includes statistical analysis of replies and retweets. At first, we define the *reply ratio* of an account m as the percentage of all replies on the account to

its all tweets (considering all tweets collected, the same below). Similarly, the *retweet ratio* of an account m is the percentage of all retweets on the account to its all tweets. Figure 4(a) and 4(b) show the CDF of two groups' reply ratio and retweet ratio, respectively.

It can be seen from Figure 5(a) that most of the CCPs' reply ratio is lower than 1%, even about 40% is 0, while legitimate accounts' reply ratio is significantly higher than CCPs', and normally distributed with an average of about 0.3. This reveals that CCPs rarely interact with others. As seen in Figure 4(b), 80% of the CCPs' retweet ratio is lower than 1%, showing that they rarely facilitate the dissemination of information, especially harmless information.

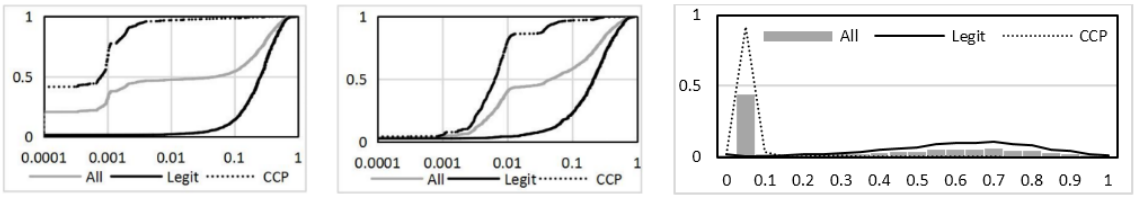


Figure 4: Distributions of reply, retweet and mention ratio, respectively. (a) CDF of reply ratio. (b) CDF of retweet ratio. (c) Histogram of mention ratio.

The above conclusions indicate that the majority of CCPs' tweets are neither replies nor retweets. Instead, they are simple tweets. The reply ratio and retweet ratio of legit group obey the normal distribution, while that of CCP group does not, and the two groups are significantly different. Moreover, legit group is loosely distributed while CCP group is more compact. So, it can be a helpful indicator for CCP detection.

3.2.2 Tweet Content Analysis. It mainly includes measures of account mention and URL usage. Similar to reply ratio, the mention ratio of an account m is defined as the percentage of all tweets with mention to its all tweets. The histogram of two groups' mention ratio is shown in Figure 4(c). From this we can see that most of the CCPs' mention ratio is lower than 0.1, while legitimate accounts' mention ratio is significantly higher than CCPs', and normally distributed with an average of about 0.7. It is also a good indicator for CCP detection.

Table 1: Online history of an example account (CCP) between April 15, 2014 and April 21, 2014.

Content	createTime
È bene fare attenzione quando tutti vi loderanno Luca evangelista	2014/4/15 10:39
http://t.co/KQY5LfaZBG	2014/4/15 14:29
Ti accorgi di toccare il fondo, quando arrivi a casa stanco e affamato, apri una bottiglia e sa di tappo!	2014/4/16 23:28
http://t.co/RZThhui7NR	2014/4/17 8:18
Hanno più foto i cuori nel caffè che io alla mia cresima.	2014/4/17 11:00
http://t.co/riNpUXaTXB	2014/4/18 7:19
Se sbaglio chiedo scusa, se ho ragione chiedo scusa comunque visto che mi porterai a sbagliare.	2014/4/18 23:38
Gli uomini sono come le salsicce:pelle fuori e maiale dentro. #sapevatelo	2014/4/19 10:49
Inutile che fate le cose di nascosto, tanto lo so che vi state organizzando, per farmi il regalo di compleanno	2014/4/19 16:10
-Voglio mostrare al mondo quello che provo per te!- ----: -Ma cosa vuoi che gliene importi al mondo!-	2014/4/19 17:09
La sensibilità è un'arma a doppio taglio : la prima lama ti apre la mente, la seconda ti squarcia il petto.	2014/4/19 22:50
Penso al senso di delusione del fondo del barile quando c'ha visto arrivare	2014/4/20 11:38
Non ogni nube porta tempesta. William Shakespeare	2014/4/20 18:48
http://t.co/G0H18Nqz2J	2014/4/21 14:29

As for URL, we observed the online history of an example account (CCP) between April 15, 2014 and April 21, 2014, in which the record is shown in Table 1. During the week, it posts 14 tweets, none of which is reply or retweet, and contains no mentions. There are 4 tweets, each of which contains only one URL, and they are basically the same format and intermingled with a few simple tweets. Considering that each of these tweets with URLs contains only one URL processed by a short URL and no other text, we can see something abnormal (legitimate users often share a link with a caption). So, we can analyze the length of all tweets with URLs (hereinafter referred to as *UTL*). Figure 5(a) shows the histogram of the two groups' UTL. Of course, a tweet contains a maximum of 140 characters. From this we can see that the UTL of legitimate accounts is evenly distributed between 20 and 130, but a large part of the population is close to 140. It is likely that one tweet is not enough to express their meaning, indicating that legitimate accounts have the intention to share as much information as possible. In contrast, nearly half of the CCPs have an UTL of 22 which is just the length of a short URL. As it turns out, CCPs send tweets with URLs without

saying another word. Thus, the accounts of the two groups are clearly differentiated in UTL.

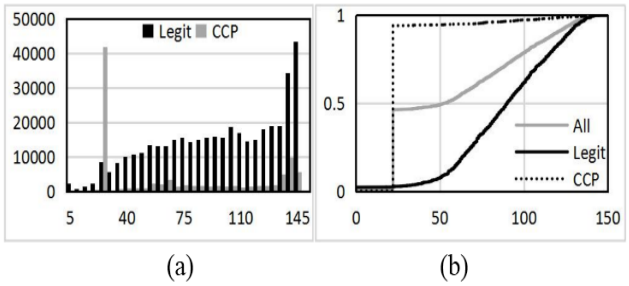


Figure 5: Analysis of URL. (a) group-wise histogram of length of tweets with URL (UTL). (b) user-wise CDF of UTL median.

Since UTL is a feature for tweets and is a distributed feature for accounts, some statistical features such as mean, median, standard deviation and so on need to be extracted. Figure 5(b) shows CDF of two groups' UTL median. More than 90% of CCPs have UTL median of 22, while the legitimate group basically presents a uniform distribution between 50 and 140. Therefore, UTL can be a good indicator for CCP detection.

3.3 Analysis of Temporal Features

Looking at Table 1, the time interval between two consecutive tweets posted by a CCP is not very short, at least three hours. Actually, legitimate accounts tend to have an online activity every few minutes [24]. So, we compute all the time intervals between two consecutive tweets of two groups, and plot the histogram in Figure 6(a), where the X-axis is in log scale of base 10 in seconds. Then, we find that both of the two groups have a bimodal distribution, but the location and size of the peaks are different. The first peak of the legit group, the main peak, is around $10^2 = 100$ seconds (several minutes), which indicates an active state. In general, a legitimate account takes about several minutes to edit and post a tweet, reply, or retweet. The second peak, also a small peak, is about $10^{4.6} \approx 40{,}000$ seconds (11 hours) which may indicate rest time. By observing the data, we find that legitimate accounts do have a segmented and continuous active state, with online behavior occurring every few minutes within each period, and the time differences between two adjacent segments are about 11 hours. For CCP group, the second peak mainly coincides with legitimate group, but the first peak is different. Its first peak is around $10^{3.9} \approx 8000$ seconds (2 hours), which is very close to its second peak, and the boundary is blurred. Therefore, CCPs don't have the so-called active state, because they are probably controlled by software. So, the time intervals of two groups are differentiated.

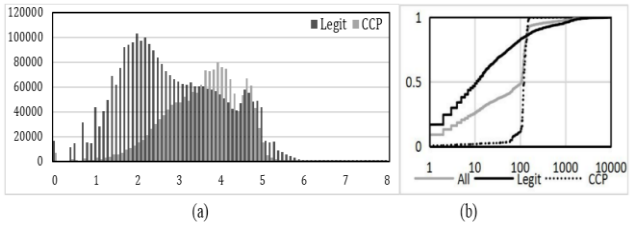


Figure 6 Analysis of temporal features, where the X-axis is in log scale. (a) group-wise histogram of time intervals between two consecutive tweets (in seconds); (b) user-wise CDF of time interval median (in minutes).

Since the time interval is a distributed feature for accounts, we will also calculate its statistical features. Figure 6(b) shows CDF of two groups' time interval median. From this we can see that most CCPs' time interval median is clustered around 100 minutes (2 hours), while legitimate group obeys the long tail distribution and is looser. So, we will select the time interval as one of the detection features.

4 The CrowdGuard Detection Method

4.1 Scenario Description and Feature Selection

As shown in section 2, our baseline in [19] only applies when historical data is sufficient, not when it is limited. It performs a detection using data up to Nov 14, two weeks after the election. By then CCPs have an adverse effect in social media environment [25]. To mitigate the negative impact, early detection is required, where CCPs can be detected at any given moment, especially on the day of the event, or even before the event. It can give early warning of catastrophic malicious events, and maintain OSN public security [26]. For example, if we want to detect at the week before the event (Oct 24), we assume that all tweets recorded by the dataset after that date are not available, only data before that date. As a result, there is no peak at the CCP group's daily activeness curve, and the amount of data available will be greatly reduced.

For an account, the number of followers, friends, age and other information can be displayed on its profile, called *explicit features*. As is shown in Section 3, reply and retweet ratio, Mention, URL, hashtag usage, time interval and so on need indirect calculation, and change dynamically over time. These features, especially those that need to be calculated from the history of tweets, are called *implicit features*. The overall scheme of our detection method, *CrowdGuard*, is that considering both of the explicit and implicit features of accounts, the mixed dataset is divided into two groups by unsupervised algorithm.

Based on the analysis in Section 3, we extract the features of account (see Table 2), and 43 features are extracted. Then we screen out the features with high degree of discrimination based on the chi-square value and the information gain [27]. As a result, the top 14 features are selected.

Table 2: List of the Features.

Type	Category	Examples	Note
Explicit	Demographics	Id, age, screen name length, description length	Age is in days, since the account created.
	Network Statistics	Followers, friends, favorites, status	Count their number and rate (subject to age)
	Boolean info	Default profile, geo enabled	
Implicit	Interaction	Activeness per day, Reply ratio, Retweet ratio	Customize the time range
	Content	URL, mention, hashtag, URL Tweet Length	Distributed features, calculating statistical features
	Temporal	Time interval between two consecutive tweets	Distributed features

4.2 Gaussian Mixture Model for Detection

As is seen in Figure 4(c), when the two groups are mixed together, they have a bimodal distribution, and each of the peaks represents a group. Not only mention ratio, but also most of the other features are similar. Therefore, Gaussian Mixture Model (GMM) [28] can best describe these mixed groups.

We need to separate two Gaussian distribution groups from mixed groups and estimate the parameters of the two groups. In this way, the problem exactly conforms to the GMM. For our problem, the probability distribution model is shown as follows.

$$P(y|\theta)=\alpha_1\varphi(y|\theta_1)+\alpha_2\varphi(y|\theta_2)\alpha_1,\alpha_2\geq 0,\alpha_1+\alpha_2=1,\theta=$$

where $\varphi(y|\theta_k)$ is the density function of the Gaussian distribution $N(\mu_k,\sigma_k^2)$, which is the k-th sub-model.

The model can be trained by EM algorithm. All observation data (accounts) need to be input to output model parameters, so as to determine which sub-model each account is more likely to belong to. In this way, all accounts are split into two groups.

5 Evaluation

5.1 Dataset and Metrics

Our original dataset is shown in Table 3.

Table 3: Statistics of dataset.

Class	Users	Tweets	Crawled at
CCP	991	1610034	2014-11-14
Legit	1083	2839361	2015-05-02

Although our method is a clustering method, since our dataset has class label, we can still use the performance measures in the standard classification method, such as precision and recall. Considering CCP as positive class and legit as negative class, our evaluation metrics include Precision, Recall, Accuracy and F_β-measure, where $\beta>0$ is a weight coefficient, measuring relative importance of precision and recall. The more β is, the more relatively important the recall is [29]. In general, β is chosen as 1, denoting that precision is as important as recall. But in our case, we should try our best to avoid undetected malicious accounts, so recall is a little more important than precision. Here we choose β as 2, and F-measure become F2-measure.

5.2 Evaluation Against Baseline Methods

5.2.1 Settings. We need to simulate an early detection scenario, and hope to detect CCPs at any time. So, two date is chosen, i.e. October 30, 2014 (while-event) and October 24, 2014 (pre-event), respectively. For each of the two detection dates, the online behavior of each account can only be selected before the detection date. Therefore, two datasets are constructed, called while-event dataset (WED) and pre-event dataset (PED).

5.2.2 Baseline Methods. **Supervised content polluters classification.** We choose the detection method proposed by Caverlee's research team [30], whose dataset is collected for about seven months through Honeypot technology they developed, including of 22223 content polluters and 19276 legitimate accounts. Using the features of User Demographics, User Friendship Networks, User Content and User History, they selected random forest classifier to classify malicious accounts, and the accuracy and F1 could exceed 98%. Now, we will use their dataset to train a random forest classifier, and test its robustness on our dataset. Since the format of the two datasets does not match, we extracted the common features between them. Nevertheless, the 10-fold cross-validation on Caverlee's dataset still achieved over 96% precision and recall.

Unsupervised sequence clustering. Different from the detection of classification method one by one, the clustering method is group detection and does not require training process. Our strong baseline, Social Fingerprinting [19] belongs to sequence clustering, which has been detailed in Section 2.2.

5.2.3 Evaluation Result. The evaluation results on WED and PED are presented in Table 4 and Table 5, respectively.

Table 4: Performance comparison between baseline and CrowdGuard on while-event dataset.

Method	Precision	Recall	Accuracy	F2
Caverlee et al. [30]	0.535	**0.971**	0.583	0.835
Social Fingerprinting [19]	0.934	0.058	0.548	0.071
CrowdGuard	0.978	0.928	0.956	0.938
CrowdGuard(select features)	**0.986**	0.947	**0.968**	**0.955**

Although Caverlee's method [30] achieves the highest recall, their precision is only about 50%, and the high false alarm rate may cause trouble to many legitimate accounts. Social Fingerprinting [19] almost has no false alarm, but recall is too low, since in the early scenario, the DNA strand was too short to make the similarity of group behavior apparent. It makes a large number of content polluters go unpunished, so it is not qualified for early detection. The reason why CrowdGuard achieves the best overall performance is that not only most of our features are robust enough, but also they are bimodal, and GMM is very fit to model them. Moreover, these features combine individual and group features, and cover many aspects, especially focusing on collective behavior, which makes content polluters more difficult to evade.

Table 5: Performance comparison between baseline and CrowdGuard on pre-event dataset.

Method	Precision	Recall	Accuracy	F2
Caverlee et al. [30]	0.535	**0.976**	0.583	0.838
Social Fingerprinting [19]	0.951	0.079	0.558	0.097
CrowdGuard	0.965	0.928	0.949	0.935
CrowdGuard(select features)	**0.969**	0.940	**0.957**	**0.946**

6 Related Work

Recently, there is many academic literature on modeling and analyzing content polluters, and many detection methods have been proposed, which can be divided into individual content polluter (ICP) detection and CCP detection.

For ICP detection, we usually adopt classification method in machine learning where the key lies in feature selection and classifier selection, and feature selection is more important. Amleshwaram et al. [31] proposed 15 features of the message content to detect content polluters in Twitter. Stringhini et al. [9] used the features of friends and message content to detect spam accounts in OSN. Thomas et al. [32] proposed a real-time URL detection scheme based on the page content to which the URL points. Egele et al. [33] extract 7 content features, modeling the messages, and judge whether the messages published later deviate from the created model to detect compromised accounts. Different features have different robustness, with some easily evaded and others difficult. Yang et al. [15] made an empirical evaluation of the robustness of 24 common features and gave possible strategies for spammers to evade detection features.

As for CCP detection, we usually detect them in a group fashion. Miller et al. [34] clustered the features of users' profile and message contents, and used StreamKM++ and DenStream combined data stream clustering algorithm to cluster legitimate accounts into one class, and those outside the class were content polluters. Wang et al. [35] used HTTP request sequences of users accessing social networks for clustering. Cao et al. [36] believed that the behaviors of content polluters showed loose synchronous behaviors in social networks, and they were detected by clustering the behaviors of accounts. Viswanath et al. [37] model the behavior of legitimate accounts through PCA, and then determine whether an account is content polluter based on the degree of deviation between itself and the model. Social Fingerprinting [19] adopts k-LCS algorithm [23] to detect CCP.

7 Conclusion and Future Work

In this paper, we focus on collective content polluters in OSN. We analyze the difference of CCPs and legitimate accounts in different aspects of features and get some findings. Using them, we extract 43 individual/group features and leverage GMM for early detection. Finally, an experimental scenario simulating early detection is designed. The results show that CrowdGuard is competent for early detection.

However, there are still some limitations. As for data, although the distinction between the two groups is obvious, there is no way that content polluters can come from only one source in the real world, and legitimate group is not necessarily made up of harmless accounts. As for methodology, group detection is really a good idea, but because there are so many users in OSN, it maybe takes a lot of time and resource to process huge amounts of data.

There are several interesting directions to explore. For model, other models can be considered, such as graph mining, anomaly detection framework [38], and so on. As for methodology, the combination of one-by-one detection and group detection can be considered. As for data, in addition to CCP, we should consider other types of content polluters and simulate real-world scenarios.

8 ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2017YFB1001800), the National Natural Science Foundation of China (No. 61772428,61725205).

REFERENCES

[1] Hongyu Gao, Jun Hu, Tuo Huang, Jingnan Wang, and Yan Chen. 2011. Security Issues in Online Social Networks. IEEE Internet Computing 15, 4 (July 2011), 56-63. DOI: http://dx.doi.org/10.1109/MIC.2011.50

[2] Ting-Kai Huang, Md Sazzadur Rahman, Harsha V. Madhyastha, Michalis Faloutsos, and Bruno Ribeiro. 2013. An analysis of socware cascades in online social networks. In Proceedings of the 22nd international conference on World Wide Web (WWW' 13). ACM, New York, NY, USA, 619-630. DOI: https://doi.org/10.1145/2488388.2488443

[3] Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Lería, Jose Lorenzo, Matei Ripeanu, Konstantin Beznosov, and Hassan Halawa. 2016. Íntegro. Comput. Secur. 61, C (August 2016), 142-168. DOI: http://dx.doi.org/10.1016/j.cose.2016.05.005

[4] Xin Ruan, Zhenyu Wu, Haining Wang, and Sushil Jajodia. 2016. Profiling Online Social Behaviors for Compromised Account Detection. IEEE Trans. Information Forensics and Security, 11(1), 176-187.

[5] Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K. Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, and Alex 'Sandy' Pentland. 2016. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. Int. J. Inf. Secur. 15, 5 (October 2016), 475-491. DOI: http://dx.doi.org/10.1007/s10207-016-0321-5

[6] Mehwish Nasim, Andrew Nguyen, Nick Lothian, Robert Cope, and LewisMitchell. 2018. Real-time Detection of Content Polluters in Partially Observable Twitter Networks. In WWW '18 Companion: The 2018 Web Conference Companion, April 23-27, 2018, Lyon, France. ACM, New York, NY, USA, 9 pages. DOI: https://doi.org/10.1145/3184558.3191574

[7] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference (ACSAC '11). ACM, New York, NY, USA, 93-102. DOI: http://dx.doi.org/10.1145/2076732.2076746

[8] Muhammad Al-Qurishi, Mabrook Al-Rakhami, Atif Alamri, Majed Alrubaian, Sk Md Mizanur Rahman, and M. Shamim Hossain. 2017. Sybil Defense Techniques in Online Social Networks: A Survey. in IEEE Access, 5, 1200-1219.

[9] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC '10). ACM, New York, NY, USA, 1-9. DOI: http://dx.doi.org/10.1145/1920261.1920263

[10] Eva Zangerle and Günther Specht. 2014. "Sorry, I was hacked": a classification of compromised twitter accounts. In Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC '14). ACM, New York, NY, USA, 587-593. DOI: https://doi.org/10.1145/2554850.2554894

[11] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. 2010. Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (IMC '10). ACM, New York, NY, USA, 35-47. DOI: http://dx.doi.org/10.1145/1879141.1879147

[12] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. Commun. ACM 59, 7 (June 2016), 96-104. DOI: https://doi.org/10.1145/2818717

[13] Tommi Vatanen, Mikael Kuusela, Eric Malmi, Tapani Raiko, Timo Aaltonen, and Yoshikazu Nagai. 2012. Semi-supervised detection of collective anomalies with an application in high energy particle physics. In IJCNN, 1-8.

[14] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI: http://dx.doi.org/10.1145/1541880.1541882

[15] Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," in IEEE Transactions on Information Forensics and Security, 8(8), 1280-1293.

[16] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. In Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 273-274. DOI: https://doi.org/10.1145/2872518.2889302

[17] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. arXiv: 1703.03107. Retrieved from https://arxiv.org/abs/1703.03107

[18] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2016. DNA-inspired online behavioral modeling and its application to spambot detection. IEEE Intelligent Systems, 31(5), 58-64.

[19] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2018. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable and Secure Computing, 15(4), 561-576.

[20] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. In Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1395-1405. DOI: https://doi.org/10.1145/2736277.2741637

[21] Tu Ngoc Nguyen, Cheng Li, and Claudia Niederée. 2017. On Early-stage Debunking Rumors on Twitter: Leveraging the Wisdom of Weak Learners. arXiv: 1709.04402. Retrieved from https://arxiv.org/abs/1703.04402

[22] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 963-972. DOI: https://doi.org/10.1145/3041021.3055135

[23] Michael Arnold and Enno Ohlebusch. 2011. Linear Time Algorithms for Generalizations of the Longest Common Substring Problem. Algorithmica 60, 4 (August 2011), 806-818. DOI: https://doi.org/10.1007/s00453-009-9369-1

[24] Huayi Li, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao. 2017. Bimodal Distribution and Co-Bursting in Review Spam Detection. In Proceedings of the 26th International Conference on World Wide Web (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1063-1072. DOI: https://doi.org/10.1145/3038912.3052582

[25] Bin Guo, Chao Chen, Daqing Zhang, Zhiwen Yu, and Alvin Chin. 2016. Mobile crowd sensing and computing: when participatory sensing meets participatory social media. in IEEE Communications Magazine, 54(2), 131-137.

[26] Bin Guo, Zhu Wang, Zhiwen Yu, Yu Wang, Neil Y. Yen, Runhe Huang, and Xingshe Zhou. 2015. Mobile Crowd Sensing and Computing: The Review of an Emerging Human-Powered Sensing Paradigm. ACM Comput. Surv. 48, 1, Article 7 (August 2015), 31 pages. DOI: https://doi.org/10.1145/2794400

[27] Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In Icml, 97, 412-420.

[28] Jeffrey A. Bilmes. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 4(510), 126.

[29] Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

[30] Kyumin Lee, Brian David, and James Caverlee. 2011. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In ICWSM, 185-192.

[31] Amit A. Amleshwaram, Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang. 2013. CATS: Characterizing automation of Twitter spammers. 2013 Fifth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, 2013, 1-10.

[32] Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, and Dawn Song. 2011. Design and Evaluation of a Real-Time URL Spam Filtering Service. In Proceedings of the 2011 IEEE Symposium on Security and Privacy (SP '11). IEEE Computer Society, Washington, DC, USA, 447-462. DOI: https://doi.org/10.1109/SP.2011.25

[33] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2015. Towards detecting compromised accounts on social networks. IEEE Transactions on Dependable & Secure Computing, 14(4), 447-460.

[34] Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. 2014. Twitter spammer detection using data stream clustering. Information Sciences, 260(1), 64-73.

[35] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y. Zhao. 2013. You are how you click: clickstream analysis for Sybil detection. In Proceedings of the 22nd USENIX conference on Security (SEC'13). USENIX Association, Berkeley, CA, USA, 241-256.

[36] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering Large Groups of Active Malicious Accounts in Online Social Networks. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14). ACM, New York, NY, USA, 477-488. DOI: https://doi.org/10.1145/2660267.2660269

[37] Bimal Viswanath, M. Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2014. Towards detecting anomalous user behavior in online social networks. In Proceedings of the 23rd USENIX conference on Security Symposium (SEC'14). USENIX Association, Berkeley, CA, USA, 223-238.

[38] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient Adaptive Exploration of Twitter Bot Networks. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17), Jana Diesner, Elena Ferrari, and Guandong Xu (Eds.). ACM, New York, NY, USA, 467-474. DOI: https://doi.org/10.1145/3110025.3110163

Supervised Lexicon Extraction for Emotion Classification

Alberto Purpura
purpuraa@dei.unipd.it
University of Padua
Padua, Italy

Chiara Masiero
chiara.masiero@statwolf.com
Statwolf
Padua, Italy

Gianmaria Silvello
silvello@dei.unipd.it
University of Padua
Padua, Italy

Gian Antonio Susto
sustogia@dei.unipd.it
University of Padua
Padua, Italy

ABSTRACT

Emotion Classification (EC) aims at assigning an emotion label to a textual document with two inputs – a set of emotion labels (e.g. anger, joy, sadness) and a document collection. The best performing approaches for EC are dictionary-based and suffer from two main limitations: (i) the out-of-vocabulary (OOV) keywords problem and (ii) they cannot be used across heterogeneous domains. In this work, we propose a way to overcome these limitations with a supervised approach based on TF-IDF indexing and Multinomial Linear Regression with Elastic-Net regularization to extract an emotion lexicon and classify short documents from diversified domains. We compare the proposed approach to state-of-the-art methods for document representation and classification by running an extensive experimental study on two shared and heterogeneous data sets.

CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; **Sentiment analysis**; • **Computing methodologies** → **Natural language processing**; **Machine learning**.

ACM Reference Format:
Alberto Purpura, Chiara Masiero, Gianmaria Silvello, and Gian Antonio Susto. 2019. Supervised Lexicon Extraction for Emotion Classification. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316700

1 INTRODUCTION

Emotion Classification (EC) [35] is a fast growing topic in text classification along with Topic Labeling (TL) [34] and Sentiment Analysis (SA) [16]. The research in the field of SA is mainly focused on detecting the subjectivity (objective or subjective) or polarity (positive or negative) of a text rather than specific emotions [20]. EC, on the other hand, is a more fine-grained SA and performs the following task: given a set of emotion labels $\mathcal{E}$ (e.g., anger, joy, sadness, etc.) and a collection of documents $\mathcal{D}$ (e.g., sentences, paragraphs or entire documents), assign to each document a label $e \in \mathcal{E}$. Developing effective techniques to detect emotions in user-generated content can be useful, for instance, to understand the position of a population towards a certain social issue or to assess the success of a marketing campaign [15]. Traditionally, EC has been performed mainly using dictionary-based approaches which employ lists of terms related to specific emotions – e.g. ANEW [3]. Nevertheless, there are two main issues limiting the broad applicability of these approaches: (i) they cannot be employed in domains where a term is used with different emotional connotations than the dictionary denotations; and, (ii) they cannot infer an emotion label for sentences that do not contain any of the known keywords (out-of-vocabulary keyword problem). To overcome these two limitations, we propose a supervised method to extract an emotion lexicon from a given textual corpus to perform EC on other documents from different domains. Our approach exploits the coefficients of a multinomial logistic regression model to extract an emotion lexicon from a collection of short textual documents. First, we extract all unigrams and bigrams in the chosen collection and consider their TF-IDF weights. Second, we train a logistic regressor to perform EC on the documents of the collection. Third, we create the emotion lexicon by considering all the unigrams and bigrams with non-zero coefficients in the logistic regressor model. We also perform an exhaustive evaluation of the proposed approach. We assess the quality of the selected terms in the lexicon as features for EC by comparing the proposed method to a popular feature selection approach: Principal Component Analysis (PCA). We evaluate the quality of the lexicon for the EC task by employing four supervised classifiers (i.e. K-Nearest Neighbors, Gaussian Naive Bayes, Support Vector Machine, Feed-Forward Neural Network) in order to assess the impact of our lexicon on different classification approaches. We evaluate the generalization power of the lexicon extraction method by employing two heterogeneous public collections (tweets and news headlines). Finally, we compare our lexicon-based document representation approach to Word2Vec [18], a widely-used method to create dense document representations; to FastText [13], a state-of-the-art method for document classification; to a set of Naive Bayes classifiers (SNBC), each trained to recognize one emotion, as done in [29]; and to the approach presented in [1], which employs a generative unigram mixture model (UMM) to model emotionality and neutrality of words from labeled documents. The main contributions of this work are:

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316700

- the application of a method based on regularized multinomial logistic regression to build an emotion lexicon;
- the feature selection process of the lexicon extraction method which employs: (i) unsupervised methods for document representation (TF-IDF and Word2Vec); and, (ii) supervised and unsupervised methods for dimensionality reduction (logistic regression and PCA);
- the evaluation of the discriminative power of the selected features by means of extensive classification experiments with four different classifiers of increasing complexity;
- the comparison of our best performing classification pipeline to #Emotional Tweets [19] – i.e., the best-known baseline on the Twitter Emotion Corpus (TEC) data set;
- the comparison of our best performing classification pipeline to SNBC [29] and UMM [1], two other recent approaches for emotion classification;
- a "transfer learning" experiment to assess the consistency of the proposed approach across different domains. We extract the lexicon from the TEC data set and we test it on the TEC itself and on the SemEval 2007 Affective Text Corpus [31] (1000 and 250 News Headlines) for the EC task;
- a shared public repository containing the open source code and references to the data to reproduce our experiments.

We show that the proposed approach performs better overall than the chosen baselines (i.e., #Emotional Tweets, UMM, SNBC) both on the homogeneous scenario (learn and test on tweets) and on the heterogeneous scenario (learn on tweets and test of news headlines). Moreover, we show that our approach for short documents representation is highly competitive with two cornerstone methods like Word2Vec and FastText.

The rest of the paper is organized as follows: in Section 2 we give an overview of the most popular strategies for EC; in Section 3 we present our approach for EC and its main components; in Section 4 we describe the shared collections used in our experiments and our experimental setup; in Section 5, we present the results of different classification systems. Finally, in Section 6 we draw some conclusions and outline future work.

2 RELATED WORK

In the last years, SA and Opinion Mining research increasingly focused on the classification of user-created contents like tweets [6, 21], news [14] or customer reviews [25] available on the Web. Despite this, the importance of a more fine-grained emotion classification (EC) on this type of documents has emerged only recently [12].

There are two main approaches to EC: the categorical one [5], which consists of assigning a label to each element to be classified; and the dimensional one [27], which attempts to represent detected emotions in a space, for example of two dimensions, such as valence (i.e. pleasure/displeasure) and arousal (i.e. activation/deactivation). In the literature, EC has been applied to different domains and textual data of different lengths, from posts on social media [35] to fairy tales [10]. Each application domain has its own peculiarities. Specifically, classification for shorter texts is usually more challenging and less effective. In this work, we focus on the analysis of short texts and on the problem of how to extract meaningful features for the classification of emotions.

Some approaches tackle emotion classification as a multi-label classification problem. In [4] for example, the authors assign zero or more emotion labels to each sentence in a movie review and propose a method to learn dependencies between labels and exploit this information during the classification. In this work, we consider the problem as a multi-class classification problem since this is the most widespread formulation, and it allows us to evaluate more easily our method of emotion lexicon extraction. This is a reasonable assumption since in our case we are dealing with short texts which usually express a single emotion.

Different sets of 6, 8 or 20 emotions [7, 11, 23, 24] have been considered in the literature; however, the set of six Ekman emotions [7] (anger, disgust, fear, joy, sadness, and surprise) has become the most popular choice in many studies such as in [32] and in SemEval 2007 "Affective Text" task [31]. For this reason, we employ this set of six emotions to evaluate our approach and compare it to #Emotional Tweets [19] where the author propose another supervised method to extract features for EC in tweets based on Pointwise Mutual Information (PMI). In [19], the author employs PMI to select word n-grams according to their correlation with emotion labels in a data set, then uses the information on the presence or absence of these features in a document to perform its classification using a Support Vector Machine (SVM) classifier.

The original approach presented in this work for document representation, dimensionality reduction and emotion lexicon extraction is based on a multinomial regression model which is employed to select the most relevant features for the classification task. To the best of our knowledge, we are the first to apply, in the EC domain, a multinomial regression model for lexicon extraction.

A related approach to ours is FastText [13], a state-of-the-art method for document classification. FastText first learns a word embedding, then averages word representations into a text representation, and finally uses softmax to compute the probability distribution on the predefined classes. This architecture is inspired to the CBOW Word2Vec [18] model. In particular, in FastText the middle word used in CBOW is replaced by a class label. This approach, similarly to what we propose in this work, uses a labeled data set to create document embeddings with discriminant features and has been successfully used for SA before [22]. For this reason, we compare our approach for EC also to FastText.

We also compare our approach to two other recent approaches for emotion classification. SNBC [29], which employs a set of Naive Bayes classifiers (SNBC) – each trained to recognize one emotion – for emotion classification; and to the approach presented in [1], which employs a generative unigram mixture model (UMM) to model emotionality and neutrality of words from labeled documents and then classify them.

3 PROPOSED APPROACH

We propose to use a labeled collection of documents where each item is associated with an emotion label and begin by indexing it with TF-IDF, extracting all the unigrams and bigrams which appear

in the collection more than 5 times. [1] Then, we train a multinomial logistic regression model with elastic-net regularization and consider its sparse coefficients matrix β (see Equation 4). For each class, we keep the features extracted with TF-IDF, which have non-zero coefficients in the respective column of β. This set of unigrams and bigrams constitutes the emotion lexicon we extract for each emotion.

The components of our lexicon extraction and evaluation pipeline are detailed in the rest of this section.

3.1 Emotion Lexicon Extraction

Document Representation. We first perform stopwords removal; [2] secondly, we create document embeddings using TF-IDF indexing [17]. TF-IDF is a well-established technique and at this step in our pipeline we index the documents considering all unigrams and bigrams in the collection.

Lexicon Extraction. The proposed approach for lexicon extraction is based on *multinomial logistic regression* (MLR) with *elastic net regularization* [2]. Logistic regression is a non-linear model for classification, also known as *logit regression, maximum-entropy (MaxEnt) or log-linear classification.* This model estimates the probabilities describing the possible outcomes of the classifier using a logistic function. For binary classification problems, suppose we have a response variable that takes values in $G = \{0, 1\}$; we denote with $y_i = I(g_i = 1)$ the indicator response variable. We model the logistic function as

$$\Pr(G = 1|X = x) = \frac{e^{\beta_0+\beta^T x}}{1 + e^{\beta_0+\beta^T x}}, \tag{1}$$

where β is a p-dimensional array and p is the size of an input sample, β_0 is a scalar and represents the intercept of the model. The objective function for the elastic net penalized logistic regression uses the negative binomial log-likelihood (where $\alpha \in [0, 1]$ is the elastic-net regularization coefficient):

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1] - \left[\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i^T\beta) - \log(1 + e^{\beta_0+x_i^T\beta})\right]. \tag{2}$$

For multi-class classification problems with K classes [30], the logistic function is

$$\Pr(G = k|X = x) = \frac{e^{\beta_{0k}+\beta_K^T x}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_\ell^T x}}, \tag{3}$$

where the response variable has K levels $\mathcal{G} = \{1, 2, ..., K\}$. Let Y be the $N \times K$ indicator response matrix, with elements $y_{i\ell} = I(g_i = \ell)$. Then, the elastic net penalized negative log-likelihood function [33] is

$$\ell(\{\beta_{0k}, \beta_k\}_1^K) = - \left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{i\ell}(\beta_{0k} + x_i^T\beta_k) - \log(\sum_{k=1}^{K} e^{\beta_{0k}+x_i^T\beta_k})\right)\right] + \lambda\left[(1-\alpha)||\beta||_F^2/2 + \alpha\sum_{j=1}^{p}||\beta||_1\right], \tag{4}$$

where β is a $(p + 1) \times K$ matrix of coefficients, β_k refers to the k-th column (for outcome category k) and β_j the j-th row (vector of K coefficients for variable j). The last penalty term is $||\beta||_1$ where we employed a lasso penalty on its coefficients in order to induce sparse solution.

To solve this optimization problem we use the partial Newton algorithm by making a partial quadratic approximation of the log-likelihood, allowing only (β_{0k}, β_k) to vary for a single class at a time. For each value of λ, we first cycle over all classes indexed by k, computing each time a partial quadratic approximation about the parameters of the current class. [3] After obtaining a model, we examine the β-coefficients for each class and keep the non-zero features extracted with TF-IDF. Finally, for EC, we consider the terms associated to non-zero coefficients and we index the documents in the experimental collections considering only these terms with their respective TF-IDF weight.

3.2 Emotion Lexicon Evaluation

In order to evaluate the quality of the features selected with our approach for lexicon extraction, we compare it to another method for feature selection and dimensionality reduction: *Principal Component Analysis* (PCA) [36]. We use PCA to decompose a multivariate data set in a set of successive orthogonal components that explain a maximum amount of the variance. In this case, our goal is to create a new lower-dimensional set of features to represent a textual document in order to later detect the emotions expressed within the document itself. We employed the randomized truncated Singular Value Decomposition (SVD) [4] to create new document embeddings of size 50 (this size leads to the best performances on both the test data sets as we have assessed empirically).

To check if the proposed approach for feature selection leads to a consistent performance improvement because of the discriminant power of the extracted lexicon, we perform EC using different classifiers of increasing complexity.

We classify the documents according to the emotions they express in a multi-class fashion. For data sets which associate more than one label to each document, we train a classifier for each class (i.e. to predict the labels "Joy" and "Not-Joy"), repeat the process for each emotion, and solve the multi-label classification problem as a set of binary classification problems. The classifiers we employ are the following:

- *K-Nearest Neighbors* (K-NN): a simple widely-known supervised classifier based on the distance between elements in

[1] The choice of this frequency threshold is based on the results of the experiments on the selected collections.

[2] We employ a list of 170 English terms, see `nltk v.3.2.5` https://www.nltk.org.

[3] A Python implementation which optimizes the parameters λ and α of the model is: https://github.com/bbalasub1/glmnet_python/blob/master/docs/glmnet_vignette.ipynb.

[4] We used the implementation available in scikit-learn library v.0.19.1 for Python http://scikit-learn.org/stable/.

the input space. Different distance metrics can be used, we considered the most widely-used Euclidean distance;
- *Gaussian Naive Bayes* (GNB): a supervised probabilistic classifier based on the Bayesian model [28], which assumes independence between the input features and does not require any hyper-parameter optimization;
- *Support Vector Machine* (SVM): one of the most popular and effective supervised classifiers [28]. This is also the classifier employed in #Emotional Tweets, the baseline to which we compare the proposed approach. In our experiments, we decided to employ an SVM with a linear kernel in order to keep the model as simple as possible and to speed up the optimization process of its hyper-parameters;
- *Feed-Forward Neural Network* (FFNN): a supervised classifier which can be of arbitrary complexity growing with the number of layers and hidden units of the network. We included this classifier in our pipeline since Neural Network and Deep Learning-based approaches are attracting great interest in the NLP and, more in general, in the Machine Learning community. The hyper-parameters that we considered for the optimization of this classifiers are the number of layers and their size. We considered different combinations of one or two layers of sizes from 5 to 5000.

In order to evaluate the generalization power of the proposed approach for lexicon extraction, we extract an emotion lexicon from the TEC data set and then perform EC on the SemEval 2007 Affective Text Corpus. The results of our evaluation are reported in Section 5.

Finally, we compare the results of different classification pipelines to #Emotional Tweets, SNBC, UMM, Word2Vec and FastText as alternative and effective methods for dense document representation.

4 EXPERIMENTAL SETUP AND REPRODUCIBILITY

The implementation of the proposed methods and the code for their evaluation is available on our public repository. [5] For the evaluation of our pipelines we employ two publicly available data sets:

- 1000 and 250 Headlines [6]: the SemEval 2007 Affective Text corpus [31] contains 1250 newspaper headlines (1000 for training and 250 for testing) labeled with the six Ekman emotions by six annotators. For each headline-emotion pair, the annotators assigned scores from 0 to 100 indicating how strongly the emotion was expressed in the headline. For our experiments, like in [19], we considered scores greater or equal to 25 to indicate that the headline expresses the corresponding emotion;
- TEC [7]: the Twitter Emotion Corpus (TEC) data set [19] contains a set of 21,051 tweets labeled with the six Ekman emotions downloaded with the Twitter API [8] and labeled according to their hashtags (i.e., #anger, #disgust, #fear, #joy, #sadness, #surprise).

We begin by training a Word2Vec model [9] to obtain vectors with 100 features by employing the skip-gram algorithm with negative sampling (5 negative samples), a window of 5 terms, and we filter out words that appear in the collection less than 10 times. We iterate the training for 5 epochs on each of the evaluation data sets. Document embeddings are obtained by averaging their word embeddings. Then, for our experiments, we first extract (for each data set) the features in an unsupervised way with TF-IDF or Word2Vec (or in a supervised way with FastText [10]); after this, we reduce the size of the TF-IDF embeddings by training the multinomial logistic regression model on the data set split into a training and a test set of equal size; finally, we apply the classifiers we described in Section 3.2. In order to statistically validate our results, we employ a 5 folds cross-validation procedure for each combination of the our proposed classification pipelines. Since the scores relative to each fold of the baseline method [19] are not available, we are not able to compute any statistical test to check if our results are statistically different from the chosen baseline; nevertheless, we report a comparison of the metrics obtained by summing and then averaging the results from each run.

The hyper-parameters of the considered classifiers have been obtained by optimizing them using the scikit-learn RandomizedSearchCV class. [11] For all our experiments we decided to keep the optimized configuration of the classifiers hyper-parameters associated to the documents embeddings created with TF-IDF on TEC data set. We also evaluated the impact of the optimization of the classifiers on the 1000 Headlines data set and noticed no relevant performance improvement with a different configuration of the hyper-parameters.

The hyper-parameters employed for each document classifier [12] are the following:

- K-NN: 2 neighbors;
- GNB: no hyper-parameters required;
- SVM: linear kernel with error penalty $C = 12.5$ on the default l_2 norm;
- FFNN: a single-layer Feed-Forward Neural Network of size 2000.

The emotion lexicon we employ has been extracted from a subset of the TEC data set (70%) randomly sampled without replacement. [13] The performance measures we employed for the evaluation of our approach on the Headlines data set displayed in Table 2 are mean precision, mean recall and mean f1 score calculated as the average of single class precision (recall, f1 score) over all considered classes. We obtained the measures relative to FastText in Table 2 converting the class probabilities returned by the algorithm to binary values according to a threshold value. We decided to evaluate FastText on the best-case scenario, selecting the threshold which led to the best results on the test data set.

[5] https://bitbucket.org/albpurpura/supervisedlexiconextractionforec/src/master/.
[6] http://web.eecs.umich.edu/~mihalcea/affectivetext/.
[7] https://bit.ly/2M0KCku.
[8] https://developer.twitter.com/en/docs/tweets/search/overview.

[9] Gensim library for Python [26].
[10] https://fasttext.cc.
[11] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.
[12] The implementation of each classifier can be found in the scikit-learn library v.0.19.1 http://scikit-learn.org/stable/.
[13] This is due to some limitations when dealing with large quantities of data of the `glmnet` library we use to train the regression model.

5 EVALUATION

5.1 Features Quality

In Figure 1 and Table 1, we report the results obtained on TEC data set. In this case, the problem to solve is single label multi-class classification. We evaluated four different classification pipelines and report the F1 score for each pipeline and emotion:

- TF-IDF: we employed TF-IDF indexing to compute the document embeddings and we classified them using four different classifiers;
- TF-IDF w/MLR: we employed a Multinomial Logistic Regressor (MLR) to extract an emotion lexicon for each class and then performed document classification using the TF-IDF embeddings relative to that lexicon;
- TF-IDF w/PCA: we employed TF-IDF indexing to compute the document embeddings and then reduced their size using PCA. We report the classification performance using four different classifiers;
- W2V: we computed document embeddings by averaging the term embeddings of each document and classified them using four different classifiers;
- ET: the #Emotional Tweets baseline on the TEC data set performance (obtained with an SVM classifier);
- MLR: we employed TF-IDF indexing to compute the document embeddings, then employed a MLR for the classification of the documents, without any feature selection step;
- SNBC: we report the results obtained on the TEC dataset by [29] with a set of Naive Bayes classifiers, each trained to predict one emotion, using unigrams and bigrams as lexical features.

Based on the charts in Figure 1 (and in Table 1 for a more accurate analysis), we can make five conclusions. First, dimensionality reduction (either with PCA or with the selected lexicon) leads generally to a performance improvement in the EC task. Second, the proposed approach for feature selection leads to a stable performance increase with almost all of the classifiers (the GNB classifier is the only exception on the TEC dataset). Third, if we compare the proposed approach for document representation to the pipeline using Word2Vec, we notice a better performance of our approach with all of the classifiers except for the GNB. Fourth, we notice that when we combine our method for lexicon extraction with the SVM classifier (or the FFNN), as it is done in #Emotional Tweets, we always obtain a better performance than the current baseline on the TEC data set. Finally, we observe that the F1 scores of the proposed approach for feature selection, when used in the same pipeline of the FFNN or SVM classifiers, are always higher or equal to the cases where only a MLR was employed for the classification, without feature selection.

We compare our approach to FastText – considered here as another state-of-the-art approach for document classification – and our method outperforms it in several cases; FastText performs better than our approach only on precision for anger and disgust and on F1 score for sadness and surprise. It is worth noticing that an advantage of our approach with respect to FastText is the *interpretability* of predictions. In fact, FastText is based on an artificial neural network and it is difficult to figure out which features play a major role in classification. On the contrary, in our pipeline we have access to the weights relative to each unigram or bigram selected by the MLR model and the SVM classifier. Hence, we can estimate the influence of each feature on the classification of each document. Moreover, our model does not require to learn an embedding for unigrams and bigrams, differently from FastText. Thus, it is more suitable for classification over small collections.

We also compare our approach to SNBC [29], and we outperform it in most of the cases with just one exception in the classification of tweets expressing anger.

Therefore, the proposed approach for lexicon extraction is a trustworthy and well-performing new method to select terms in a corpus which are discriminative for the classification and it outperforms the current baselines.

5.2 Generalization Power and Overall Performance

In Table 2 we compare our approach for emotion lexicon extraction with the method proposed in #Emotional Tweets, a Word2Vec-based document representation, FastText, and the UMM approach presented in [1]. We extract an emotion lexicon (unigrams and bigrams) from TEC data set then, we evaluate the classification accuracy with the lexicon extracted from the training data set on the 250 Headlines collection. [14] We employed this experimental setup to be able to compare our approach to the chosen baseline system [19] under the same conditions. Contrarily to [19] – for the hardware limitations of our setup – we employ only 70% of TEC data set for lexicon extraction instead of the whole collection, then perform the evaluation on the 250 Headlines collection. For the experiments using Word2Vec and FastText, we trained, respectively, the GNB classifier and the FastText model, on the 1000 Headlines collection.

As we can see in Table 2, our method for the extraction of an emotion lexicon outperforms the other approaches and the baseline in terms of recall and F1 score.

Moreover, we observe that the number of features in the emotion lexicon extracted with our approach is roughly half of those of the baseline; this indicates that we are selecting a more restricted subset of elements, without compromising the overall classification quality.

In the comparison with UMM, we considered its best performing case, where additional information was used and training and testing were performed on data from the same domain. The authors employed the following additional features: Part-Of-Speech (POS) tags and Contextual Features (CF) which include punctuation marks, emoticons, capitalized words, elongated words, negations and sentiment features. [15] Even though in [1] the authors indicate only the average F1 score of their evaluation results, we see that our approach – even in a cross-domain classification scenario – outperforms UMM.

[14] We used the 1000 Headlines collection as a training set for the classifier as done in the baseline system.

[15] For a complete description of these features, please refer to [1]

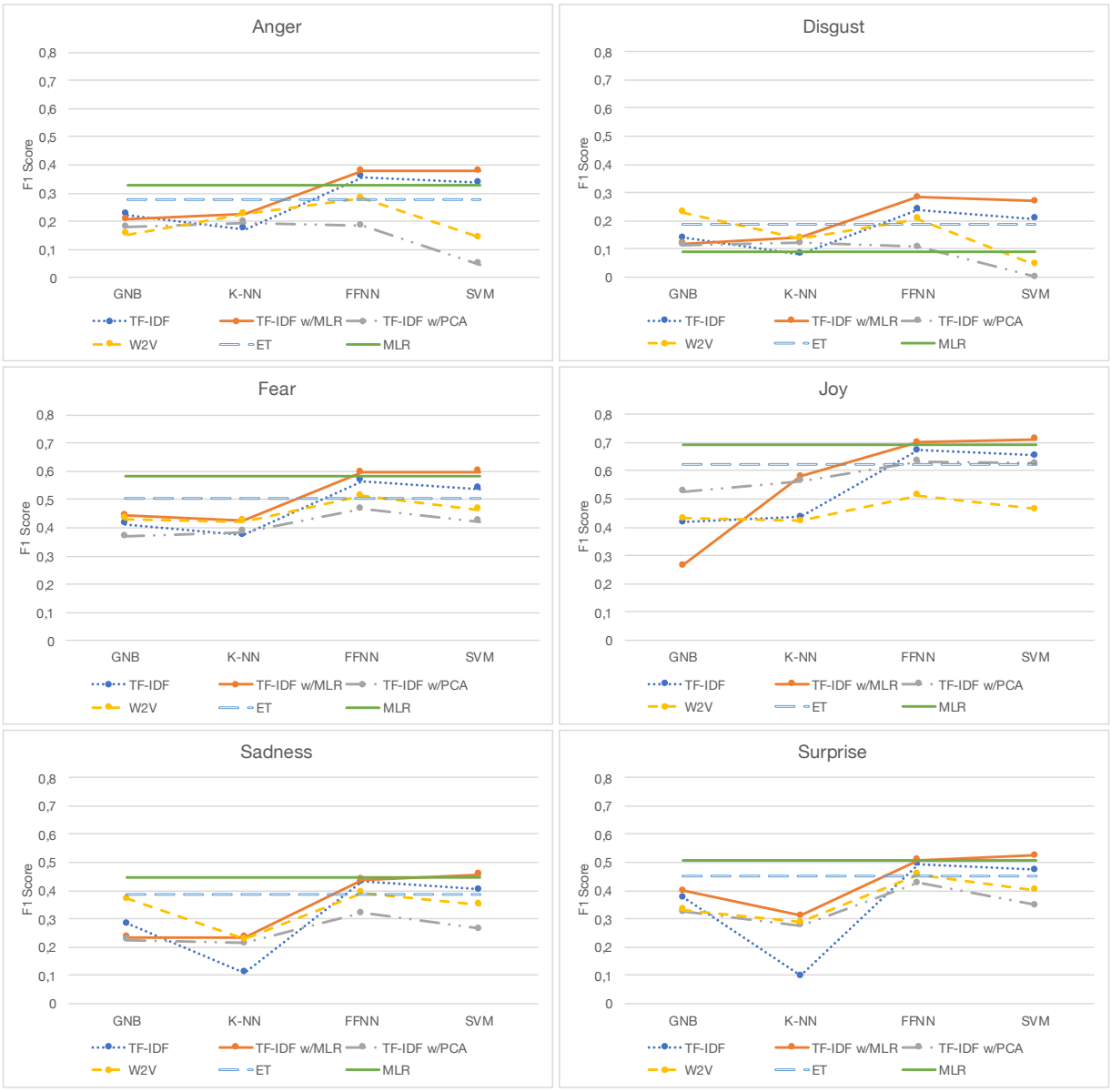


Figure 1: Classification results on TEC. The F1 scores were computed considering the sum of the total number of documents belonging to the class, the total number of correct predictions and the total number of performed predictions for each class.

6 CONCLUSIONS AND FUTURE WORK

In this work, we proposed and evaluated a novel approach for emotion lexicon extraction. The process employs a multinomial logistic regressor to extract an emotion lexicon from a labeled collection of documents.

To evaluate the quality of the extracted lexicon, we performed two different tests: (i) we considered lexicon extraction as a feature selection/dimensionality reduction problem and compared it to PCA; (ii) we compared the classification performance of different classifiers with the proposed lexicon-based document representation in order to assess the discriminant power of the lexicon terms.

The results of the above experiments showed a consistent performance improvement in the EC task when employing the proposed approach. We also compared our approach for document representation to Word2Vec, considered as the most widespread alternative to obtain dense document representations, to FastText, a state-of-the-art method for document classification, and to SNBC [29]. With respect to Word2Vec and SNBC, we obtained better performances for almost all emotions (with the exception of the emotion anger, where SNBC is the best option), and we are competitive also with FastText, which outperforms our approach only in a handful of

cases. Furthermore, the proposed approach for EC has the advantage of being more interpretable than embedding-based ones since the features we used to perform the classification are easily accessible and represented by word unigrams and bigrams. Finally, we evaluated the generalization power of the lexicon extraction process by generating a lexicon from TEC data set, and using that for the classification of the documents in the SemEval 2007 Affective Text corpus. Also in this case, our approach performed better overall than #Emotional Tweets [19] and UMM [1] the baseline systems chosen as reference.

All the information necessary to reproduce our experiments, including the details about the training of the classifiers, is provided. We also make our code publicly available.

We highlight that our approach might be applied to document classification also for other tasks, such as topic labeling or sentiment analysis. Indeed, we are using a general approach adaptable to any task or applicative domain in the document classification field. Another element of investigation will be the analysis of the impact of the pre-processing step for document representation. Finally, we plan to conduct a thorough statistical analysis by means of general linear mixed models in order to determine the contribution of each

Table 1: Classification results on TEC data set (full results in Figure 1). We highlight in bold the top performance value of each measure for each emotion.

Emotion	Doc. Repr.	Dim. Red.	Classifier	Prec.	Rec.	F1 Score	Emotion	Doc. Repr.	Dim. Red.	Classifier	Prec.	Rec.	F1 Score
anger	TF-IDF	MLR	GNB	0.137	0.428	0.208	**joy**	TF-IDF	MLR	GNB	**0.734**	0.161	0.264
			K-NN	0.179	0.309	0.226				K-NN	0.532	0.637	0.580
			FFNN	0.395	0.365	**0.379**				FFNN	0.655	0.751	0.700
			SVM	0.418	0.344	0.377				SVM	0.657	**0.777**	**0.712**
		none	GNB	0.162	0.356	0.222			none	GNB	0.618	0.316	0.418
			K-NN	0.148	0.209	0.173				K-NN	0.526	0.374	0.437
			FFNN	0.369	0.348	0.358				FFNN	0.659	0.687	0.672
			SVM	0.339	0.333	0.336				SVM	0.642	0.666	0.654
		PCA	GNB	0.113	0.440	0.180			PCA	GNB	0.591	0.474	0.526
			K-NN	0.147	0.291	0.195				K-NN	0.542	0.586	0.563
			FFNN	0.247	0.148	0.185				FFNN	0.556	0.735	0.633
			SVM	0.328	0.025	0.047				SVM	0.479	0.899	0.625
	W2V	none	FFNN	0.312	0.254	0.280		W2V	none	FFNN	0.636	0.681	0.658
	ET	none	SVM	0.373	0.223	0.279		ET	none	SVM	0.645	0.604	0.624
	FastText	none	none	**0.422**	0.233	0.300		FastText	none	none	0.654	0.765	0.705
	SNBC	none	none	0.304	**0.452**	0.363		SNBC	none	none	0.72	0.691	0.705
disgust	TF-IDF	MLR	GNB	0.067	0.522	0.118	**sadness**	TF-IDF	MLR	GNB	0.332	0.181	0.235
			K-NN	0.131	0.152	0.141				K-NN	0.328	0.183	0.235
			FFNN	0.327	0.251	**0.284**				FFNN	0.457	0.422	0.439
			SVM	0.300	0.244	0.270				SVM	**0.484**	0.435	0.458
		none	GNB	0.097	0.259	0.141			none	GNB	0.277	0.290	0.283
			K-NN	0.045	**0.530**	0.082				K-NN	0.388	0.063	0.109
			FFNN	0.274	0.213	0.239				FFNN	0.437	0.425	0.431
			SVM	0.217	0.200	0.208				SVM	0.412	0.395	0.403
		PCA	GNB	0.076	0.230	0.114			PCA	GNB	0.332	0.170	0.225
			K-NN	0.092	0.179	0.121				K-NN	0.296	0.169	0.215
			FFNN	0.170	0.079	0.108				FFNN	0.356	0.292	0.321
			SVM	0.000	0.000	0.000				SVM	0.378	0.205	0.266
	W2V	none	FFNN	0.250	0.176	0.207		W2V	none	FFNN	0.402	0.385	0.393
	ET	none	SVM	0.307	0.134	0.187		ET	none	SVM	0.419	0.360	0.387
	FastText	none	none	**0.350**	0.200	0.255		FastText	none	none	0.467	0.510	0.488
	SNBC	none	none	0.171	0.427	0.243		SNBC	none	none	0.470	**0.517**	**0.492**
fear	TF-IDF	MLR	GNB	0.391	0.508	0.442	**surprise**	TF-IDF	MLR	GNB	0.485	0.339	0.399
			K-NN	0.350	0.537	0.424				K-NN	0.641	0.205	0.311
			FFNN	0.619	**0.573**	0.595				FFNN	0.538	0.480	0.508
			SVM	**0.642**	0.562	**0.599**				SVM	0.551	0.500	0.524
		none	GNB	0.349	0.509	0.414			none	GNB	0.411	0.348	0.377
			K-NN	0.358	0.389	0.373				K-NN	**0.706**	0.053	0.099
			FFNN	0.569	0.562	0.565				FFNN	0.491	0.497	0.494
			SVM	0.543	0.536	0.539				SVM	0.475	0.473	0.474
		PCA	GNB	0.489	0.298	0.370			PCA	GNB	0.423	0.265	0.326
			K-NN	0.323	0.476	0.385				K-NN	0.568	0.184	0.278
			FFNN	0.511	0.430	0.467				FFNN	0.457	0.400	0.426
			SVM	**0.642**	0.314	0.422				SVM	0.511	0.265	0.349
	W2V	none	FFNN	0.536	0.490	0.512		W2V	none	FFNN	0.439	0.478	0.458
	ET	none	SVM	0.596	0.439	0.506		ET	none	SVM	0.506	0.405	0.450
	FastText	none	none	0.569	0.510	0.488		FastText	none	none	0.561	**0.502**	**0.530**
	SNBC	none	none	0.634	0.503	0.561		SNBC	none	none	0.626	0.403	0.489

Table 2: Comparison with #Emotional Tweets and UMM [1] (best pipeline on the dataset) on the task of lexicon extraction from the TEC data set, evaluated on 250 Headlines data set, considering the best performing pipelines.

Method	# of Features extracted from TEC lexicon	Mean Precision	Mean Recall	Mean F1 Score
#Emotional Tweets	11,418	**0.444**	0.353	0.393
TF-IDF+MLR+GNB	6,383	0.377	**0.790**	**0.479**
Word2Vec + GNB	100 (document embeddings size)	0.309	0.423	0.346
FastText	100 (document embeddings size)	0.442	0.509	0.378
UMM (ngrams + POS + CF)	-	-	-	0.410

component to the overall pipeline similarly to what has been done for off-the-shelf information retrieval systems in [8, 9].

REFERENCES

[1] Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. 2017. Lexicon based feature extraction for emotion text classification. *Pattern recognition letters* 93 (2017), 133–142.

[2] Christopher Bishop. 2016. *Pattern recognition and machine learning*. Springer-Verlag, New York.

[3] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Citeseer.

[4] Lars Buitinck, Jesse Van Amerongen, Ed Tan, and Maarten de Rijke. 2015. Multi-emotion detection in user-generated reviews. In *Proc. of the 37th European Conference on IR Research, ECIR 2015 (LNCS)*, Vol. 9022. Springer, 43–48.

[5] Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence* 29, 3 (2013), 527–543.

[6] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 519–535.

[7] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[8] Nicola Ferro and Gianmaria Silvello. 2016. A General Linear Mixed Models Approach to Study System Component Effects. In *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM Press, New York, USA.

[9] Nicola Ferro and Gianmaria Silvello. 2017. Towards an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)* 69, 2 (2017), 187–200.

[10] Virginia Francisco and Pablo Gervás. 2006. Automated mark up of affective information in english texts. In *Proc. of the 9th International Conference on Text, Speech and Dialogue, TSD 2006 (LNCS)*, Vol. 4188. Springer, 375–382.

[11] Nico H Frijda. 2017. *The laws of emotion*. Psychology Press.

[12] Mohammed Jabreel and Antonio Moreno. 2018. EiTAKA at SemEval-2018 Task 1: An Ensemble of N-Channels ConvNet and XGboost Regressors for Emotion Analysis of Tweets. *arXiv preprint arXiv:1802.09233* (2018).

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. (2016). arXiv:1607.01759 http://arxiv.org/abs/1607.01759

[14] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification.. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, Vol. 333. AAAI Press, 2267–2273.

[15] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[16] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011)*. IEEE Computer Society, 81–88.

[17] James Mayfield and Paul McNamee. 1998. Indexing using both n-grams and words. In *Proc. of the 7th Text REtrieval Conference, TREC 1998*, Vol. Special Publication 500-242. National Institute of Standards and Technology (NIST), 361–365.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*. 3111–3119.

[19] Saif M Mohammad. 2012. # Emotional tweets. In *Proc. of the First Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 246–255.

[20] Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing* 5, 2 (2014), 101–111.

[21] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1–18.

[22] Leonard Papenmeier and Christoph Friedrich. 2017. Fasttext and Gradient Boosted Trees at GermEval-2017 on Relevance Classification and Document-level Polarity. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*. 30–35.

[23] W Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. Psychology Press.

[24] Robert Plutchik and Henry Kellerman (Eds.). 1980. *Theories of Emotion (Volume 1): Theory, Research, and Experience*. New York: Academic Press.

[25] Alberto Purpura, Chiara Masiero, and Gian Antonio Susto. 2018. WS4ABSA: an NMF-based Weakly-Supervised Approach for Aspect-Based Sentiment Analysis with Application to Online Reviews. In *Proc. of the 24th International Syposium on Methodologies for Intelligent Systems (ISMIS 2018)*. Springer.

[26] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, 45–50.

[27] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[28] Stuart J Russell and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited.

[29] Ameneh Gholipour Shahraki and Osmar R Zaiane. 2017. Lexical and learning-based emotion mining from text. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*.

[30] Noah Simon, Jerome Friedman, and Trevor Hastie. 2013. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529* (2013).

[31] Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. Association for Computational Linguistics, 70–74.

[32] Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proc. of the 2008 ACM Symposium on Applied Computing (SAC)*. ACM Press, 1556–1560.

[33] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. 2012. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, 2 (2012), 245–266.

[34] Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 90–94.

[35] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing Twitter "Big Data" for Automatic Emotion Identification. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Conference on Social Computing, SocialCom 2012*. IEEE Computer Society, 587–592.

[36] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

Leveraging Motifs to Model the Temporal Dynamics of Diffusion Networks

Soumajyoti Sarkar
Arizona State University
Tempe, AZ
ssarka18@asu.edu

Hamidreza Alvari
Arizona State University
Tempe, AZ
halvari@asu.edu

Paulo Shakarian
Arizona State University
Tempe, AZ
shak@asu.edu

ABSTRACT

Information diffusion mechanisms based on social influence models are mainly studied using likelihood of adoption when active neighbors expose a user to a message. The problem arises primarily from the fact that for the most part, this explicit information of who-exposed-whom among a group of active neighbors in a social network, before a susceptible node is infected is not available. In this paper, we attempt to understand the diffusion process through information cascades by studying the temporal network structure of the cascades. In doing so, we accommodate the effect of exposures from active neighbors of a node through a network pruning technique that leverages network motifs to identify potential infectors responsible for exposures from among those active neighbors. We attempt to evaluate the effectiveness of the components used in modeling cascade dynamics and especially whether the additional effect of the exposure information is useful. Following this model, we develop an inference algorithm namely InferCut, that uses parameters learned from the model and the exposure information to predict the actual parent node of each potentially susceptible user in a given cascade. Empirical evaluation on a real world dataset from Weibo social network demonstrate the significance of incorporating exposure information in recovering the exact parents of the exposed users at the early stages of the diffusion process.

CCS CONCEPTS

• **Human-centered computing** → **Social networks**; **Social media**; **Social network analysis**.

KEYWORDS

Social Networks; Information Cascades; Network Motifs

ACM Reference Format:
Soumajyoti Sarkar, Hamidreza Alvari, and Paulo Shakarian. 2019. Leveraging Motifs to Model the Temporal Dynamics of Diffusion Networks. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316703

1 INTRODUCTION

Information, genuine or spurious, spreads over time. Modeling information diffusion to understand how information propagates among a population of individuals has been a very popular research problem in the social network and machine learning community. With availability of real world data, these models can be manipulated to overcome several challenges [1]. Most models in information diffusion work on the premise of influence from active neighbors: the effect of multiple active neighbors sequentially or collectively in an arbitrary order selectively enables the spread of information. These models implicitly assume the effect of multiple exposures from these active neighbors. However it does not account for how well a neighbor is structurally or temporally placed in the social network with respect to the current contagion that could eventually decide the role of a particular neighbor in influencing a user [7].

In this paper, we try to address the notion of "impactful" exposures in the overall framework of influence and the diffusion mechanism where we would elaborate what we mean by "impactful" in the technical section - briefly, these are exposures from among a *subset* of active neighbors of a user that are structurally and temporally positioned to influence that user in an ongoing contagion. To this end, we study the spread of information by understanding the evolution of information cascades[8] which are simply messages reshared by multiple individuals in a social network. To test whether the information obtained from these "impactful" exposures as opposed to all active neighbors is useful, we use the exposure information in a system that models these information cascades. However, for evaluating how this model with these exposures compares to the existing methods, we devise a prediction problem that infers the parents (the user which ultimately infects a user) of a susceptible user and test these models on this prediction problem. This prediction problem is similar to the link prediction problem [11] although we consider the temporal setting within a more constrained setup. Summarizing, the main two less studied shortcomings of the existing information diffusion models that we focus on are as follows:

(1) The majority of the models that consider either social influence or viral propagation processes through active neighbors cannot identify neighbors who were responsible for exposure or consider all active neighbors of a susceptible user as nodes which exposed that user. In fact, one of the popular ways of assessing the impact of exposures is to consider the collective influence of *all active neighbors* of a susceptible user and use it in a threshold model [6, 9].
(2) Secondly, network inference methods [4, 5] that attempt to recover edges in a static network or in a temporal setting to predict links ahead of time assume complete observations from the start of diffusion (of the cascade) till a certain time. We attempt at inferring the activation sources in a constant

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316703

evolving manner by using temporal networks and the current available information instead of all information from the beginning.

Studying the effect of filtered exposures facilitates understanding the diffusion process and thereafter inferring the parent nodes in the following way: (1) first this allows us to test the hypothesis that all exposures from active neighbors are not responsible in the collective influence of an infected node and since exposures precede infection, this paves the way toward inferring the parent node of an infected user from among the active neighbors by the process of elimination, (2) secondly we leverage network motifs [12] which are recurring subgraphs in a network occurring significantly, for identifying exposure nodes for a susceptible user with respect to information cascades and the social network information (the past historical data). The advantage of using motifs as would be elaborated further is that it takes advantage of the position of the neighbors within the cascade as well as the past historical social network to assess whether an active neighbor of a susceptible user can even expose a user to the message and therefore be a potential candidate for its parent. Following this, our main contributions are as follows:

- We devise a method that uses network motifs to prune out the potential exposure nodes for a susceptible user at different stages of an information cascade.
- We use this additional information from the exposure nodes to model the evolution of the information cascade based on techniques from survival analysis introduced in [4].
- To test the effectiveness of this exposure information and therefore the model of diffusion, we conduct a suit of experiments on a real world dataset from Weibo to predict the parent (from whom users reshare a piece of message) of a user. We make the following observations: (1) taking exposure into account helps in the parent inference process in the early stages of the cascade compared to the models with just temporal information (2) the parent inference using the cascade topology becomes difficult as the cascade progresses, suggesting that the structure loses any specific organization thus failing to use any additional information on motifs.

The rest of the paper is as follows: we first give a brief description of the technical preliminaries on information cascades and diffusion in Section 2, followed by the model of information diffusion demonstrating the exposure effect in Section 4, the dataset description in Section5.0.1 and finally the results discussion in Section 5.

2 TECHNICAL PRELIMINARIES

We shall represent historical social network information by a directed graph $G_{\mathcal{H}} = (V_{\mathcal{H}}, E_{\mathcal{H}})$. In this graph, $V_{\mathcal{H}}$ denotes the individuals who were either involved in historical diffusion processes (thus building historical diffusion networks) or are part of the follower-followee network. We denote by $e = (i, j) \in E_{\mathcal{H}}$ to imply that an information can be propagated from user i to user j given that a similar diffusion took place in history or by virtue of the follower-followee network. These relationships are known apriori using observed or available data. Further, an event or a message can originate from node $u \in V_{\mathcal{H}}$ and can be reshared by a neighbor of u in the network $G_{\mathcal{H}}$. This produces a cascade C of reshares which

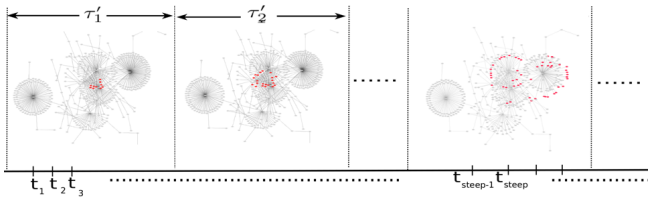


Figure 1: The cascade partitioning procedure. The likelihood for the cascades is computed over this incrementally evolving network. The number of nodes activated in each subsequence τ' is kept fixed. Red nodes denote new activated nodes in each subsequence.

comprises individuals $V_C = \{v_1, v_2, \ldots, v_{\mathcal{R}}\}$ ordered by their corresponding chronologically sorted infection times $\tau_C = \langle t_1, t_2 \ldots t_{\mathcal{R}} \rangle$, where $\mathcal{R}$ denotes the number of reshares/infections of C or size of the cascade. We refer to these individuals as adopters of C. In addition, we assume that ultimately each individual $v \in V_C$ is infected by exactly one existing adopter of C which we refer to as its *parent* in the cascade lifecycle. This assumption relies on the observation that a user can reply to or retweet one piece of information multiple times, however it posts it the first time being activated at a single point in time. We denote the underlying network produced by the participants of a cascade C by $G_C = (V_C, E_C)$ where an edge $e = (u, v) \in E_C$ denotes that either v reshared C from an adopter u which we term $parent(v)$ or there is an edge e in $E_{\mathcal{H}}$. In simple terms, we add the social/historical diffusion network links from $G_{\mathcal{H}}$ to create an augmented cascade network. We will use the set $Parents(V')$ to denote the set of parents for each node in $V' \in V_C$. We denote the time at which node v reshared a cascade C as $t_{v,C}$. Also, we will drop the subscript C from the notations when an operation described on a cascade is applicable for all cascades. We note that the presence of social/historical network edges introduces cycles in the structure of G_C which otherwise would exhibit a tree structure inherent to the property of cascades.

We use the notation τ' to denote a subsequence of τ. For the purposes of notation, we will consider the subsequence τ'' as the succeeding subsequence of τ' for any two subsequences τ' and τ'' of τ. In our work, we create the sequence of subsequences $\tau = \langle \tau'_1, \ldots, \tau'_Q \rangle$, ordered by the starting time of each subsequence, where we denote Q to be the number of subsequences for C, which would vary for different cascades. The method of splitting the cascade C into a sequence of such subsequences τ' is described in the following section.

Temporal network evolution

In context of G^{τ}, we denote $G^{\tau'} = (V^{\tau'}, E^{\tau'})$ as the subgraph of G^{τ}, where $V^{\tau'}$ denotes the individuals who reshared a cascade C in the time subsequence τ' and $E^{\tau'}$ denotes the set of edges $e = (u, v)$ where the resharing from u to v happened in the time subsequence τ' or there was an interaction in the historical diffusion period indicated by the presence of e in E_D. For the subsequences, the following conditions hold: (1) $|V^{\tau'_i}| = |V^{\tau'_j}|$ and (2) $E^{\tau'_i} \cap E^{\tau'_j} = \emptyset$, $\forall i \neq j, \in [0, Q]$. We note that the condition $|\tau'_i| \neq |\tau'_j|$ may or may

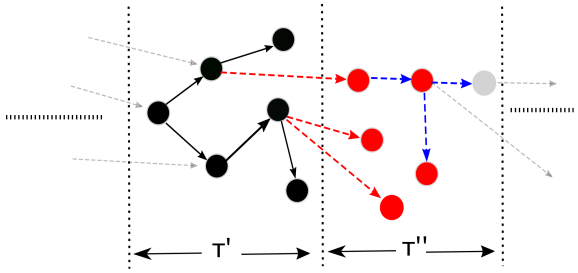


Figure 2: The problem statement setting. Black nodes represent the observed portion of the cascade in a subsequence $\tau' \in \tau_C$ in terms of diffusion. Red nodes denote the sample of nodes with known attributes which participates in τ'' but with unknown *parent* nodes. Blue edges are those which are ignored in the reconstruction process. Observe that the potential parents form the ends of cut edges between $V^{\tau'}$ and $V^{\tau''}$. We infer these red edges.

not hold for any $i \neq j$, that is to say the time range spanned by the subsequences in itself may differ depending on the time taken by $G^{\tau'}$ to form the network. In our work, we select and keep $|V^{\tau'}|$ fixed for every cascade in our corpus. Since we analyze the subsequences in a sequence that depicts the evolution of the cascade over time, the advantage of selecting this subsequence node set size a-priori is that we can avoid retrospective observation of the entire cascade lifecycle and incrementally progress with the network analysis using motifs over intervals until we reach N_{inhib}. This gives us the freedom to be agnostic about the final cascade size $|V^{\tau}|$ or T_C for selecting the subsequence size $|V^{\tau'}|$ or span $|\tau'|$.

A *temporal representation* of a cascade is denoted by a sequence of overlapping subsequences $\mathcal{N} = \langle N_1, \ldots, N_S \rangle$ such that the following conditions hold: $V^{N_i} = V^{\tau'_{i-1}} \cup V^{\tau'_i}$, and $E^{N_i} = E^{\tau'_{i-1}} \cup E^{\tau'_i}$ $\forall\ i \in [1, Q]$. We perform network analysis on each N where we drop the index subscript when we generalize the analysis for all subsequences for all cascades. Such a temporal representation $\mathcal{N}$ helps us in avoiding disjoint subsequences for network analysis and replicates a sliding window approach. This method to partition the cascade into subsequences would help us in two ways: (i) having the network topology of any subsequence $G^{\tau'}$, we can attempt to find $parent(v)$, $v \in V^{\tau''}$ from among nodes in V^{τ}, where τ'' denotes the subsequences immediately exceeding τ', and (ii) we can prune the nodes in $G^{\tau'}$ to compute the *exposure nodes* temporally relevant to each node in $G^{\tau''}$, for which we use network motifs that would be described later. In this way, instead of considering all information from the start of the cascade, we only consider the recent past to predict the parents of future nodes. An example of the cascade network procedure that governs the network evolution is shown in Figure 1.

3 PROBLEM STATEMENT

Given a temporal network in the current time phase of the cascade lifecycle and a list of susceptible and *to-be* infected nodes in the next timestamp (Figure 2), our goal is to discover the set of nodes from this current temporal network that are going to activate these susceptible nodes. The problem is formally defined as follows: *Given a graph* $G_C^{\tau'} = (V^{\tau'}, E^{\tau'}), V^{\tau''}$, *such that* $\tau', \tau'' \in \tau_C$, τ'' *succeeds* τ' *in the set of subsequences, our goal is to find* $parent(v)$ $\forall v \in V^{\tau''}$ *with the condition* $parent(v) \in V^{\tau'}$. The constraint that the set of parents for nodes in $V^{\tau''}$ must be part of $V^{\tau'}$ would not always hold as shown in Figure 2. Note that we do not predict whether the susceptible nodes would be infected but as a retrospective measure, we attempt to find their parents. Our problem statement in that respect introduces two new directions for research:

(1) It sets this problem apart from general network inference where network edges are inferred irrespective of whether there is a chance of having a node as a parent of another node through diffusion mechanism governed by network topology.
(2) A single parent inference procedure ultimately paves the way towards diffusion path reconstruction in cascades since as the stages (or subsequences) unfold over time, each node can be traced back to its parent and the diffusion tree can be reconstructed. It is not necessary to have a fully connected tree as considered in [14] to be able to construct the tree in a forward progressing manner.

We do note that this temporal parcellation of the cascade introduces different lags for parent consideration for different nodes (since the subsequences are not compact and bounded manually by *time* but by node size of subsequences). However empirically, we found that the time difference between a node's activation time and it's parent's activation time follows a power law where most users reshare a message from parents who reshared the same message not far before. Therefore using information from only the recent past to predict the parent generalizes well in the real-world scenario.

4 MODEL LEARNING AND INFERENCE

In this section, we describe in details the 3 components that lay the roadmap: finding the potential exposure nodes using motifs to parametrically modeling the diffusion process to finally extending the learned model parameters and the motifs to infer the parents of future nodes in the cascade lifecycle. To model the likelihood of diffusion between the set of nodes in a subsequence τ'', we would use the information from the preceding subsequence τ'. We proceed through the stages $[1, Q_C]$ to aggregate the likelihood for C. We summarize the 3 components as follows: (1) we extract the parent node (available from data) and prune the exposure nodes for each node v that participated in the cascade C prior to its own infection, (2) we use the parent and exposure nodes for all participants to compute the likelihood of the cascades relying on previous techniques used in survival analysis and formulate a convex objective function and we use an optimization algorithm for learning the parameters used in the likelihood function and (3) finally, given a test subsequence of nodes in τ' and the sample of nodes from τ'', we devise an inference algorithm based on the learned model parameters to infer $Parents(V^{\tau''}) \in V^{\tau'}$. We note that the parametric model helps in the prediction problem on unseen cascades and the network motifs are used for pruning out exposure nodes. We combine these two to infer the parent of a susceptible node - without the model and just the information about exposure nodes,

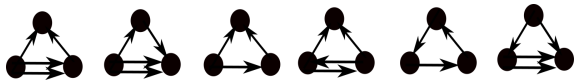


Figure 3: Motif patterns used in computing the exposure nodes. Note that some of the patterns contain parallel nodes - one of them denoting a social / historical diffusion link apart from the cascade diffusion link. These motifs have been chosen based on significance in the frequencies obtained in a few selected cascades - there are many ways of assessing significance, like z-scores, count. We use count (these appear more frequently than 1000 other motif patterns) from a few observed cascades to select these patterns.

we would have no mechanism to ultimately select one node among the filtered exposure nodes.

4.1 Extracting parent and exposure nodes

As a first step in building the model as part of the training stage, we iteratively take two consecutive subsequences τ', τ'' in a cascade lifecycle in order until the end of the cascade lifecycle. Furthermore, assuming that we know
$Parents(V^{\tau''})$ (as ground truth labels from the data) during each stage τ'', our main goal in this step is to infer for each node $v \in V^{\tau''}$, the set of exposure nodes $exp(v) \in V^{\tau'} \setminus parent(v)$. In order to prune the nodes which are relevant to v and which could have potentially exposed v to C, we use network motifs to extract subgraphs in the network $G^{\tau''} \cup G^{\tau'}$. A network motif $M = (V_M, E_M)$ is defined as an induced subgraph of G, such that $\forall e \in E_M$, it is always that $e \in V_G$. To prune the exposure nodes $exp(v)$ for a node v using network motifs, we consider the following three steps: (1) We first extract all the instances of 3-size motifs containing v from $G^{\tau''} \cup G^{\tau'}$, whose patterns match one of the patterns shown in Figure 3, (2) following this, for each $v \in V^{\tau''}$, we prune those motif instances among all such that each instance contains the $parent(v)$ among the other two nodes apart from v with the remaining third vertex $v' \in V^{\tau'}$. We collect the list $\mathcal{L}$ containing the set $\{v'\}$ from all pruned instances to form an initial set of potential exposure nodes for v, (3) finally following the work of [12], which uses motif nodes as input elements in a gating function to control the output of another node, we use these potential exposure nodes $\mathcal{L}$ together with $parent(v)$ as input to a message passing switch controlling signals to the output vertex v. For that we first multiply the product of time differences $\frac{1}{t_v - t_w}$ $\forall w \in \mathcal{L}$ (Note that $t_v - t_w$ always > 0). We start removing nodes from the list of the sorted time differences such that the final product of the time differences crosses a threshold. We denote this remaining list of nodes as $exp(v)$. Essentially, this last step helps prune out nodes which do not satisfy the AND condition for a positive output in a gate where instead of boolean values we consider real values in a continuous space. We present a detailed algorithm for this step in the Appendix section. Apart from the gating mechanism that motifs are typically used for, the fact that the network motifs appear at frequencies much higher than expected at random [16], suggests that the frequent appearance of a node v along with its potential exposure nodes in certain patterns is a result of its frequent interactions with those nodes in the past and a possible exposure effect from among other peers. At the end of this step, we have a set of nodes $C = \{v_1, v_2, \ldots, v_{\mathcal{R}}\}$ and the exposure nodes $\zeta_C = \{\exp(v_1), \ldots, exp(v_{\mathcal{R}})\}$.

4.2 Model of diffusion

In order to model the cascade likelihood, we borrow concepts from survival analysis which has been widely used for computing the diffusion likelihood of the cascades [4]. We first briefly describe a few concepts before formulating the likelihood objective function.
Survival Analysis. Let T be a non-negative random variable denoting the time when an event happens or an individual is infected. Let $f(t)$ be the probability density function of T for which $F(t) = Pr(T \leq t) = \int_0^T f(x)\, dx$ is the its cumulative distribution function. The survival function $S(t)$ gives the probability that an event does not happen up to time t, $S(t) = Pr(T \leq t) = 1 - F(t) = \int_t^\infty f(x)\, dx$. Thus, $S(t)$ is a continuous and monotonically decreasing function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. Given $f(t)$ and $S(t)$, the hazard function $h(t)$ is the instantaneous rate that an event will happen within an infinitesimally small interval just after time t given it has not happened yet up to time t, $h(t) = \lim_{\Delta t\to\infty} \frac{Pr(t \leq T \leq t+\Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$.
Computing the likelihood of cascades. The activations for a cascade are represented by the sequence of tuples (j, i, t_i) where $j, i \in E_C$, j is the parent of i and t_i is the timestamp at which i was activated. Here the propagation happened from j to i denoting that $j = parent(i)$ in the diffusion tree. The sequence of tuples are chronologically ordered by their activation times. All nodes $k \in V_C$ such that $t_k \leq t_i$ and $k \neq j$, are termed as *non-parent* nodes of i. We note that $exp(v)$ is a subset of the set of non-parent nodes since the exposure nodes implicitly satisfy the constraints of non-parent nodes.

Each directed edge in the cascade network G_C, $j \to i$ is associated with a transmission function $f_{ji}(t_i|t_j)$, which is the conditional likelihood of the infection being spread to node i at time t_i given that the same infection has already affected node j at time t_j. These functions are parametrized by parameters α_{ji} for every pair although later we will show how to remove the static nature of these pairwise unique transmission rates. We focus on shift-invariant transmission functions whose value only depends on the time difference, i.e. , $f_{ji}(t_i|t_j) = f_{ji}(t_i - t_j) = f_{ji}(\Delta_{ji})$ where $\Delta_{ji} := t_i - t_j$. If there is no infection edge from j to i, we will have $f_{ji}(\Delta_{ji}) = 0$ and $h_{ji}(\Delta_{ji}) = 0$. For each directed diffusion link from $k \to i$, such that $k \in \zeta_{i,C}$, we associate an exposure function E_{ki} parametrized by η_k that quantifies the extent to which the exposure node k affects the node i in adopting the cascade C. An illustration of various node components in the diffusion mechanism is shown in Figure 4.

The likelihood of the observed infections of a cascade C can be formulated as $f(\mathbf{t}; \mathbf{A}, \mathbf{Z}) = \prod_{t_i \leq T_{\mathcal{R}}} f(t_i|t_1, \ldots, t_{\mathcal{R}}; \mathbf{A}, \mathbf{Z})$, where A denotes the set of transmission parameters for each pair $u, i, \in E_C$, and Z denotes the set of exposure parameters for each node $w \in \zeta_C$. Since we assume node infections are conditionally independent given the parents of the infected nodes, the likelihood factorizes over the nodes. The likelihood that the parent j of an individual i infected it at time t, $t \in [t_1, T_{\mathcal{R}}]$ and that i survived from all other

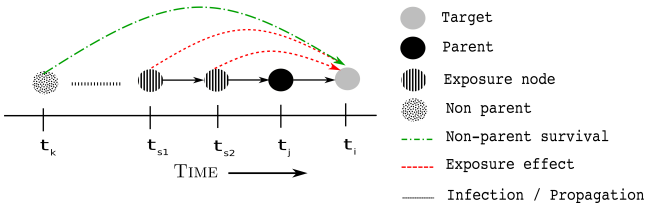


Figure 4: The components used for computing likelihood of node activations in a cascade.

adopters of C prior to time t is given by

$$TL = f(t_i|t_j;\alpha_{j,i}) \times \prod_{j\neq k;\ t_k<t_i} S(t_i|t_k;\alpha_{k,i})$$

To add the exposure effect on node i from nodes in $exp(i)$, we have the exposure likelihood term $EL = \prod_{m\in\zeta_i} E(t_i|t_m;\eta_m)$. Then the likelihood of an individual cascade is defined as follows:

$$\mathbf{f}(\mathbf{t}^{\leq T_{\mathcal{R}}};\mathbf{A},\mathbf{Z}) = \prod_{t_i\leq T_{\mathcal{R}}} f(t_i|t_j;\alpha_{j,i}) \times \prod_{j\neq k;\ t_k<t_i} S(t_i|t_k;\alpha_{k,i}) \times \prod_{m\in\zeta_i} E(t_i|t_m;\eta_m)$$

Note that the exposure term $E(.)$ is computed from only the exposure nodes extracted from the motifs on the subsequences - so the likelihood is computed subsequence wise in order of time. Finally, the likelihood of all the cascades is defined as $\prod_{c\in C} \mathbf{f}(\mathbf{t^c};\mathbf{A},\mathbf{Z})$.

Centrality modulated Rayleigh model. In this paper, we adopt the Rayleigh model for the transmission and the survival functions. Given a pair of nodes j and i, the density, hazard and survival functions of Rayleigh distributions are: $f_{ji}(\Delta_{ji}) = \alpha_{ji} \,.\, \Delta_{ji} \,.\, exp\Big(-\frac{1}{2} \,.\, \alpha_{ji} \,.\, \Delta_{ji}^2\Big)$, $h_{ji}(\Delta_{ji}) = \alpha_{ji} \,.\, \Delta_{ji}$ and $S_{ji}(\Delta_{ji}) = exp\Big(-\frac{1}{2} \,.\, \alpha_{ji} \,.\, \Delta_{ji}^2\Big)$ respectively. The drawback of directly learning the node specific and pairwise the above parameters $\mathbf{A}$, $\mathbf{Z}$ is that, firstly, for any pair (j,i) of nodes in the test data for which the infection likelihood is calculated and not observed in training data, it would not be possible in any way to compute or infer them instantly. Secondly, following positive results in [2], we use the centrality of nodes from the historical diffusion/social network to decompose these parameters in terms of node attributes, so that the corresponding transmission or exposure parameters of any pair of nodes during test time can be inferred instantaneously, if the node attributes are known. We decompose the parameters as $\alpha_{ji} = x_i \,.\, x_j \,.\, \lambda$, where the symbol x denotes the node attribute computed from the diffusion knowledge network and λ denotes the corresponding scaling control parameter. The exposure function on a node j is defined as $E_{j,m} = \eta_m \,.\, exp(-(t_j - t_m))$ and the exposure parameter is defined accordingly as $\eta_m = x_m \,.\, \beta$.

4.2.1 Objective Formulation. Considering N distinct nodes, the likelihood optimization is formulated as follows:

$$\begin{aligned} \underset{A,Z}{\text{minimize}} \quad & -\sum_{c\in C} \log f(t^c;A,Z) \\ \text{subject to} \quad & \alpha_{j,i} \geq 0, \forall i = 1\ldots N, i\neq j \\ & \eta_m \geq 0, \forall m = 1\ldots N \end{aligned} \tag{1}$$

Algorithm 1: Inferring the cut edges: InferCut

Input: $G^{\tau'}$, $V^{\tau''}$, τ', τ'', X the node attributes, λ, β
Output: E_{cut}, the cut edges between $V^{\tau'}$ and $V^{\tau''}$
$E_{cut} \leftarrow \phi$
for *each node* v *in* $V^{\tau''}$ **do**
  $parent(v) \leftarrow$ NULL, max_LL $\leftarrow$ -INF
  for *each pair* $(s,e) \in E^{\tau'}$ **do**
    /* Remove temporally far nodes */
    if $t_{v,C} - t_{s,C} > t_{thresh}$ **then**
      continue
    end
    $\alpha \leftarrow \lambda$ * $X[v]$ * $X[e]$, $\eta \leftarrow \beta$ * $X[s]$
    Calculate Transmission LL TL using α, s, v
    Calculate Exposure LL EL using η, e
    if EL * $TL > max_LL$ **then**
      max_LL $\leftarrow$ EL * TL
      $parent(v) \leftarrow$ s
    end
  end
  $E_{cut} \leftarrow E_{cut} \cup (parent(v), v)$
end
return E_{cut}

We restrict the proof here due to shortage of space but it can be decomposed into a sum of convex functions.

Considering the centrality modulated Rayleigh model, the final objective function to be optimized can be written as

$$G(\mathbf{A},\mathbf{Z},\lambda,\beta) = G_1(\mathbf{A},\mathbf{Z}) + nG_2(\lambda) + mG_3(\beta) \tag{2}$$

$$G_1(\mathbf{A},\mathbf{Z}) = -\log L(\mathbf{A},\mathbf{Z}) \tag{3}$$

$$G_2(\lambda) = \frac{1}{2P}\|\alpha - X\circ X \,.\, \lambda\| + \gamma_\lambda\|\lambda\|_1 \tag{4}$$

$$G_3(\beta) = \frac{1}{2P}\|\eta - X \,.\, \beta\| + \gamma_\beta\|\beta\|_1 \tag{5}$$

where P denotes the number of unique pairs of infections in the cascades. We prove that the final objective function in Equation 2 has a global minimum which we have included in the Appendix Section. Following this, we use standard Lasso solvers for Equation 4 and Equation 5 and the gradient descent with co-ordinate descent strategy method for solving Equation 3.

4.3 Parent Inference

Finally for the prediction problem, we consider a test subsequence τ'' in the cascade lifecycle. For inferring the set $Parents(V^{\tau''})$ for a subsequence τ'', we consider nodes in $G^{\tau'}$ in the preceding interval τ', whose timestamps, node attributes and edges are taken as input as shown in Figure 2, and another set of nodes $V^{\tau''}$ with known node attributes (or sampled from distribution) whose timestamps are known. Our main task is to infer $Parents(V^{\tau''})$ with $v' \in V^{\tau'}, \forall v' \in Parents(V^{\tau''})$ and return the cut-edges E_{cut} between these two sets of nodes. The main difficulty of this case of dynamic network inference is that as an inference task, we are not predicting the diffusion edges that might have occurred within $V^{\tau''}$, an example of which is shown in Figure 2 marked by blue edges. Due to this condition, the number of actual cut edges is low since many nodes

in $V^{\tau''}$ would be activated because of the influence of nodes within $V^{\tau''}$ itself or due to nodes activated prior to τ' and not due to nodes in $V^{\tau'}$. We present the algorithm for inference in Algorithm 1. For inferring $parent(v)$, $v \in V^{\tau''}$, we form 3-size motifs using s, e, v, the edge $(s, e) \in E^{\tau'}$. That is, for each such directed edge, we consider s as potential parent and e as exposure node and calculate the product of likelihoods. We take that pair (s, e) for v which has the highest product and consider the corresponding s as the parent.

5 EXPERIMENTAL EVALUATION

In this section, we first describe the statistics of the dataset used for learning and evaluation. We perform a suite of experiments to evaluate our model with respect to some state-of-the art methods that are able to handle these kind of problems.

5.0.1 Dataset. To build the historical diffusion network, we use the dataset provided by WISE 2012 Challenge[1]. The dataset contains user data and the reposting information of each microblog along with the reposting times which enables us to form the cascades for each microblog separately. The diffusion network $G_{\mathcal{H}}$ mentioned in the *Preliminaries Section* is created by linking any two users who are involved in a microblog reposting action within the period May 1, 2011 and July 31, 2011. The primary purpose of $\mathcal{H}$ is two fold: acting as a social network to exact possible diffusion paths in future and for computing the node attributes x used in the model. This historical diffusion network has 6,470,135 nodes and 58,308,645 edges where the average node degree is 18.02. From the corpus of cascades which spanned between June 1, 2011 and August 31, 2011, we only work with cascades with more than 300 nodes in order to have sufficient information to evaluate the temporal networks in an incremental fashion. We collect a total of 4000 cascades in the process and we train using 3000 samples and put 1000 cascades aside for testing.

Distribution of historical interactions. As a first experiment, we measure the following metric A_{v2u} as has been defined in [6] as the number of times u reshared a message from v irrespective of the message and it happened in history. For each user u in the training set that participated in some cascade in the training set, we compute the distributions of A_{v2u} for two sets of v: one where v in $exp(u)$ for the cascade in consideration and the other v in $nbrs(u)$ $exp(u)$ where $nbrs(u)$ denote the neighbors of a user u in $G_{\mathcal{H}}$. Figure 5 shows the density distributions of this measure for the two sets. It reveals that *the nodes which motifs prune as exposure nodes are also the ones who are responsible for more propagation compared to the nodes which are just neighbors but had been pruned out by motifs.*

Baseline Methods. For evaluating the performance of the inference algorithm, let E_{cut} be the set of actual edges between $V^{\tau'}$ and $V^{\tau''}$ and $\hat{E}_{cut}$ be the set of edges inferred by our algorithm. We define the following metrics of the classification model as follows:

$Precision = \frac{1}{|C|} \sum_{c \in C} \frac{|E_{cut,c} \cap \hat{E}_{cut,c}|}{|\hat{E}_{cut,c}|}$ and

$Recall = \frac{1}{|C|} \sum_{c \in C} \frac{|E_{cut,c} \cap \hat{E}_{cut,c}|}{|E_{cut,c}|}$. Simply put, we measure what fraction of the predicted edges are part of the ground truth edges as precision, and what fraction of the ground truth edges are part of the predicted edges as recall. Since for each node in $V^{\tau''}$, we try to predict and evaluate its parent in $V^{\tau'}$ only when the ground truth parent is in $V^{\tau'}$ while ignoring the rest of the nodes in $V^{\tau''}$, we evaluate the performance of the models on recall. So, for every node in $V^{\tau''}$, every model gets only one chance to select one node among the nodes in $V^{\tau'}$ and therefore the number of predicted edges is equal to the number of ground truth edges making the precision and recall same for evaluation unlike a network inference problem. Since there are no direct methods which have the same setting as our problem, we compare our framework with some recent advances in the field of information diffusion models that predict cascade adoption in the presence of active neighbors and exposures. We replicate one method that looks at all active neighbors without any pruning techniques [6] to model the probability of action adoption using the measure A_{v2u} and a second method that uses a probabilistic version of the General Threshold [9] but taking exposures into account. Additionally we consider variations of our own model. The following frameworks have been used for comparison: (1) **InferCut - No Exposure(NE)**: For this procedure, we follow the same InferCut algorithm except that in Algorithm 1, we replace the likelihood value by just the transmission likelihood and avoid the other factor of exposure likelihood. We perform this to examine whether the additional effect of exposures have any influence on adoption, (2) **Bernoulli Model**: We use this model to find the most probable nodes in $V^{\tau'}$ that infects the nodes in $V^{\tau''}$. Under this model, the probability that a user u influences a node v is given by the MLE estimator of success probability which comes from a Bernoulli trial. It is computed as $\frac{A_{v2u}}{A_v}$, A_v denoting the number of previous cascades v has participated in, (3) **Complex Contagion (CC) Model**: This model [3] is a variation of the General Threshold Model where the likelihood of adoption increases with more active neighbors. To approximate threshold-like behavior, they used a version of the logistic sigmoid function to calculate join probability of exposures, and (4) **Random - Temporal**: For each node in $V^{\tau''}$, we randomly select a node in $V^{\tau'}$ as its parent node that is Δt time units within its own occurrence. For temporal heuristics for consideration, we rule out nodes in $V^{\tau'}$ which are temporally far away from it.

For both the Bernoulli Model and the CC model, we compute the measures for all possible parents for each target node and only consider the node among the parents which has the highest value in terms of that measure.

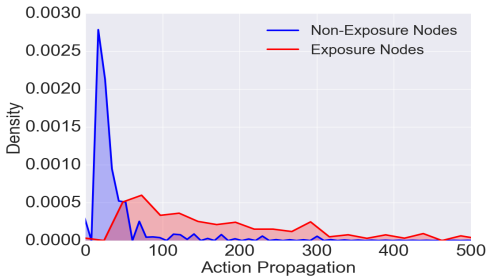
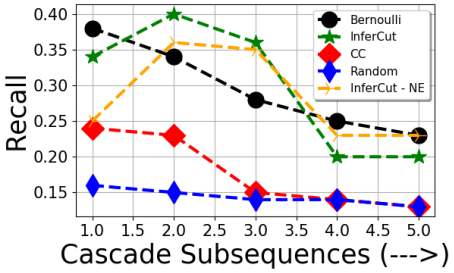
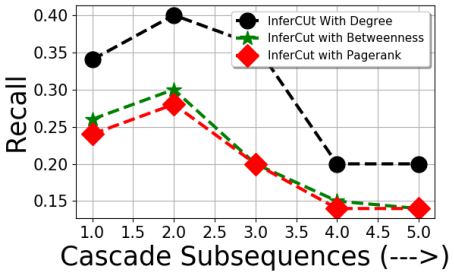


Figure 5: Distribution of A_{v2u} - the propagation count from user v to user u.

[1] http://www.wise2012.cs.ucy.ac.cy/challenge.html



(a)

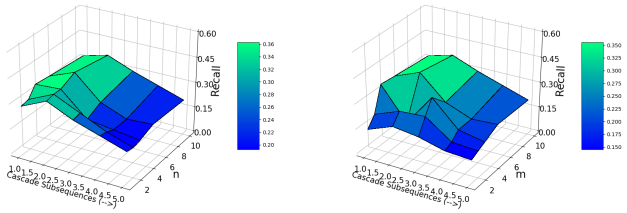


(b)

Figure 6: (a) Recall over the time subsequences from the start of the cascade, (b) Comparison of different centralities for the Rayleigh model.

Experimental setup and evaluation metrics. For our cascade splitting method described in the Preliminaries Section, we fix $|V^{\tau'}|$ to 40 for all the cascades. That is to say, we consider a set of 40 nodes as parents when trying to predict the cut edges with the other endpoints being the susceptible target nodes in the succeeding subsequence. We use the FANMOD algorithm in [18] which is a widely used algorithm to find network motifs. For training we accumulate 3000 cascades and set aside an additional 1000 cascades for testing. These cascades have been selected by shuffling through the set randomly, yet we only perform the training one time due to the expensive nature of the computation. We set the parameters $m = 5$ and $n = 5$ in Equation 2 for all the evaluation results shown although we show a hyperparameter sensitivity towards the end. We evaluate our models by considering progressing intervals/subsequences of nodes from the start for all the cascades leaving the first interval. For an interval $i \in [1, Q_C]$ we consider the network induced in the preceding interval $i-1$ as the input to InferCut and for each node v in $V^{\tau'_i}$, we predict $parent(v)$. As shown in Figure 6, we consider 5 intervals from start for every cascade, and since we consider cascades over 300 reshares, these intervals are guaranteed to exist for all cascades in our dataset.

Observations. We present the results for all the models shown in Figure 6 from which we find that the low recall of the Random - Temporal approach suggests the challenging nature of this problem. The exposure effect with InferCut helps in obtaining a maximum recall in the initial stages of the cascades recovering



(a) (b)

Figure 7: Sensitivity of the model w.r.t (a) n that controls the effect of the transmission between node pairs, (b) m that controls the effect of exposures.

almost 40% of the actual edges in the second interval, although after the third interval, the performance degrades below the No Exposure - InferCut method. One reason for this performance is that over time, the topology of the cascade structure changes rapidly to an extent that it cannot be well represented by network motifs making the exposure effect play a very little role in the diffusion towards the end. We find that the sigmoid function in the CC model performs the worst among all the models apart from the random case, and one of the reasons being is that the model heavily relies on just the exposure count and not the time difference nor the node attributes - in a way just the exposure count information is not sufficient for diffusion processes. In addition, we experiment with different node attributes for the centrality modulated Rayleigh model where the transmission rate is parameterized by the node attributes of the edge nodes. We test with 3 centralities which are pre-computed based on $G_{\mathcal{H}}$: nodal degree, betweenness and Pagerank. From Figure 6(b), we find that while the model performs the best with nodal degree, the Pagerank measure yields poor performance compared to the other two. This shows that a users' decision is largely impacted by the local network surrounding it and therefore the degree plays a larger role than measures which consider the global network structure.

Hyperparameter Sensitivity. We discuss the model sensitivity towards the hyperparameters n and m which control how much the transmission and the exposure respectively impact the performance of prediction. From Figure 7(b), we find that the performance in terms of recall gradually increases till the value of 5 after which it saturates, denoting that lowering the exposure hyperparameter that uses the node attribute values reduces the number of correctly recovered edges and thus establishing the importance of the node attributes in terms of the exposure hyperparameter. However in Figure 7(a), we find that the transmission hyperparameter does not have much impact when varied over the ranges of m.

6 RELATED WORK

Modeling the dynamics of network diffusion [4, 15] has attracted widespread attention recently due to its applications in spread of epidemics. Modeling the trajectory of the cascade growth has been studied in [2], where the authors build a model to understand the underlying diffusion process and in the event predict the future course of the cascade lifecycle. The closest in the area of reconstructing epidemics over time has been done in [14] where the authors

use Steiner trees to infer the propagation structure. However in most real datasets, observing the cascade from the start is difficult due to time of observation [17] and missing data issues. In that context, our work only relies on the current information available to determine the source of infections. Similarly, a recent study [10] on recovering the stochastic temporal networks of diffusion rely on the null models of dyadic mutually independent interactions. Although external influence or cumulative exposures [13] has been attributed to information diffusion in microblogging platforms, the absence of any concrete study of how exposures affect the diffusion rate has been one of motivations of this work.

7 CONCLUSION

We evaluated the need and effectiveness of additional information available from the partial cascade structure, in determining source of infections for future nodes. We further showed how effective this information over time could be across different stages of the cascade lifecycle, while being able to prune relevant nodes for this additional information to keep the computation scalable. One of the directions in which this research can be extended is to be able to rank exposure nodes in terms of certain functions in order to be able to directly infer the source of infection with high certainty.

Acknowledgements. Some of the authors are supported through the ARO grant W911NF-15-1-0282.

REFERENCES

[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.
[2] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, and S. Yang. Cascading outbreak prediction in networks: A data-driven approach. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 901–909, New York, NY, USA, 2013. ACM.
[3] C. Fink, A. Schmidt, V. Barash, J. Kelly, C. Cameron, and M. Macy. Investigating the observability of complex contagion in empirical social networks. In *ICWSM*, pages 121–130, 2016.
[4] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 561–568, 2011.
[5] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, New York, NY, USA, 2010.
[6] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.
[7] Z. Katona, P. P. Zubcsek, and M. Sarvary. Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research*, 48(3):425–443, 2011.
[8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, New York, NY, USA, 2003. ACM.
[9] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.
[10] X. Li and X. Li. Reconstruction of stochastic temporal networks through diffusive arrival times. *Nature Communications*, 8, 06 2017.
[11] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, May 2007.
[12] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
[13] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 33–41, 2012.
[14] P. Rozenshtein, A. Gionis, B. A. Prakash, and J. Vreeken. Reconstructing an epidemic over time. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1835–1844, New York, NY, USA, 2016. ACM.
[15] S. Sarkar, R. Guo, and P. Shakarian. Understanding and forecasting lifecycle events in information cascades. *Social Network Analysis and Mining*, 7(1):55, 2017.
[16] M. R. M. S. A. U. Shen-Orr, Shai S. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Reviews Genetics*, 31:64–68, April 2002.
[17] S. Sundareisan, J. Vreeken, and B. A. Prakash. Hidden hazards: Finding missing nodes in large graph epidemics. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 415–423. SIAM, 2015.
[18] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

COMPUTE THE EXPOSURE NODES OF A NODE v IN C

Algorithm 2: Computing $\zeta_{v,C}$

Input: Verties $V^{\tau''}$, Pre-selected Motif patterns $\mathcal{M}_{\mathcal{E}}$, temporal network G^N, threshold th for the AND gate constraint, size k of motifs.
Output: $\zeta_{v,C}$ for all $v \in V^{\tau''}$

$\mathcal{M}_C^N$, $\{M_C^N\}$ = compute_motifs $(G^N, \text{size}=k)$; // M denotes all patterns found in G^N, {M} denotes the list of 3-vertex maps for the corresponding patterns
for *each pattern* $M \in \mathcal{M}_C^N$ **do**
  if *M is not ismorphic to any pattern in* $\mathcal{M}_E$ **then**
    Skip this pattern and continue
  end
  for *each tgt in* $V^{\tau''}$ **do**
    $src \leftarrow parent(tgt)$;
    $\zeta_{tgt,C} \leftarrow \phi$;
    $\delta_{tgt,C} \leftarrow \{\}$; // empty hash table data structure
    for *each* $m \in \{M_C^N\}$ **do**
      if *src or tgt is not part of vertices in m* **then**
        Skip this instance and continue
      end
      $tv = V(m) \setminus \{src, tgt\}$; // third vertex in m
      if $t_{tv} < t_{tgt}$ **then**
        $\zeta_{tgt,C} \leftarrow \zeta_{tgt,C} \cup tv$;
        $\delta_{tgt,C}[tv] \leftarrow \frac{1}{t_{tgt} - t_{tv}}$;
      end
    end
  end
  /* AND gating constraint: iterate over all the elements of map until their product crosses the threshold */
  prod $\leftarrow$ multiply $(\delta_{tgt,C}.values())$;
  sorted_map $\leftarrow$ sort $\delta_{tgt,C}$ by its elements ;
  while $prod \leq th$ **do**
    Delete the first element of sorted_map ;
    prod $\leftarrow$ multiply $(sorted_map.values())$;
  end
end
$\zeta_{v,C} \leftarrow sorted_map.keys()$;
return $\zeta_{v,C}$

Learning Influence Probabilities and Modelling Influence Diffusion in Twitter

Zizhu Zhang, Weiliang Zhao and Jian Yang
Macquarie University
Sydney, Australia
zizhu.zhang@hdr.mq.edu.au,{weiliang.zhao,jian.yang}@mq.edu.au

Cecile Paris and Surya Nepal
CSIRO
Sydney, Australia
{cecile.paris,surya.nepal}@data61.csiro.au

ABSTRACT

Influence diffusion has been widely studied in social networks for applications such as service promotion and marketing. There are two challenging issues here: (1) how we measure people's influence on others; (2) how we predict whom would be influenced by a particular person and when people would be influenced. Existing works have not captured the temporal and structural characteristics of influence diffusion in Twitter. In this paper, we firstly develop a model to learn influence probabilities between users in Twitter from their action history; secondly, we introduce diffusion models that are used to predict how information is propagated in Twitter. Experiment results show that our proposed models outperform existing models in terms of the balanced precision and recall.

KEYWORDS

influence probability; influence diffusion; time decay; depth decay; Twitter

ACM Reference Format:
Zizhu Zhang, Weiliang Zhao and Jian Yang and Cecile Paris and Surya Nepal. 2019. Learning Influence Probabilities and Modelling Influence Diffusion in Twitter. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316701

1 INTRODUCTION

People's influence in social media can be measured by others' reactions to their postings [16]. There has been increasing interest in influence diffusion in social networks due to its applications in viral marketing [11], protest event detection [1], rumour control [6], etc. Two interesting and yet challenging issues arise: (1) how we measure people's influence on others; and (2) how we predict whom would be influenced by a particular person and when people would be influenced.

Following Kempe et al.'s seminal work in 2003 [11], simple heuristic influence probability assignments (e.g., uniform models and degree-weighted models) are being used [5]. These works are not concerned with influence probability calculation between two users. A few studies [10, 12, 15] have developed methods in learning influence probabilities from users' historical action data in Twitter. Designed for information sharing, Twitter is an ideal social network platform to investigate influence diffusion due to its *explicit interaction* mechanisms such as "retweet", "like", "reply" and "quote". For example, we can say intuitively that user v is influenced by user u if v retweets u's message [9, 12]. Twitter users can also be influenced implicitly via the *follow* mechanism, with which users are exposed to the messages posted by whom they follow. As a matter of fact there is a high possibility that a user posts a new message without any explicit interactions with other messages (a message we call it a "tweet") based on the messages of whom she or he follows. We refer to this case as an *implicit interaction*. An example is given in Table 1: user $u3391$ posts "Marriage Bill has passed", and 139 minutes later, $u849$, who follows $u3391$, posts a "tweet" of similar content. Existing works mainly use some of the explicit interactions to get the estimated influence probabilities between users. The "like" and implicit interactions are largely ignored. In our *Darwin* dataset (to be introduced in Section 6.1), the number of "likes" is larger than the sum of all the other explicit interactions, and 64% more interactions can be captured if implicit interactions are considered. It may lead to inaccurate estimations by ignoring them. Covering all of v's reactions to u's messages, explicitly and implicitly, is crucial to accurately measure u's influence on v in Twitter.

Now we address the second issue: how do we predict whom will be influenced by a particular person at what time? To answer this question, we need to look at how influence diffuses. In Twitter, the influence can change with time and depth when going through the network. The well-known diffusion models in [11] use unchanged influence probabilities, which are not suitable for Twitter.

In order to overcome the limitations of the existing works mentioned above, we propose an approach for influence probability calculation and influence diffusion modelling in Twitter. We collect two Twitter datasets, referred to as *Darwin* and *MelCup17*, to evaluate the performance of proposed models. The contributions of this paper are summarised as follows.

- To learn the pair-wise influence probabilities in Twitter, we utilise all the available explicit interactions, including the ignored "like" in existing works, and study the implicit interactions via the *follow* mechanism;
- Time and depth-sensitive diffusion models are proposed for Twitter to predict whom would be influenced by a particular user and when people would be influenced;
- Experimental results against two real-world datasets show that our proposed diffusion models outperform existing ones in terms of the balanced precision and recall, especially to predict when people will be influenced.

The reminder of the paper is organised as follows. Section 2 reviews the related work. Section 3 provides the specifications of influence in Twitter. Section 4 presents our methods for influence

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316701

Table 1: Example of implicit influence

uid	time	text
$u3391$	t-139min	RT: The **Marriage Bill** has **passed** the #Senate
$u849$	t	Big news! Same-sex **marriage bill passes** Senate
$u849$ *follows* $u3391$.		
$u849$*'s tweet is posted 139 minutes later than* $u3391$*'s message.*		
The two messages have similar content.		

probability calculation and influence diffusion modelling. We discuss the evaluation metrics in Section 5 . The experimental results are given and analysed in Section 6. We conclude in Section 7.

2 RELATED WORK

Existing work mainly use three kinds of heuristics to assign influence probabilities: uniform, trivalency and degree-weighted [5]. Saito et al. [19] are the first to focus on learning edge probabilities from diffusion logs. The probabilities for all the edges are iteratively computed using Expectation Maximization algorithm. Their method suffers the scalability due to that it updates the influence probability in each iteration. Goyal et al. [9] propose a method of learning influence probability from users' action logs sorted by action types and time. In the evaluation, they scan each actual action record and adopt an instance of General Threshold Model to update $p_v(S)$, which is the probability that v will take a particular action considering a set of his neighbours S having taken the action. Then $p_v(S)$ is compared with a randomly chosen threshold θ_v to do the prediction. While their way of evaluation is elegant, it is not able to simulate who will influence whom given just the user(s) who first take an action. Kutzkov et al. [12] learn $p(u, v)$ in Twitter data streams. They only study the "retweet probability" by adopting the ratio idea in [9]. Mei et al. [15] cover more explicit interactions, such as "reply", "mention" and "like", in the probability calculation. Jendoubi et al. [10] use belief functions to combine multiple aspects of influence indicators, including "retweet", "mention", and the "number of followers", to estimate influence in Twitter. None of these studies consider the implicit influence and fail to catch u's actions that have influenced v in a high degree of accuracy.

Another area related to this study is analysing influence diffusion in social networks. A number of well-known models are discussed in [11]. Modelling the spread of information and influence maximization have been studied in [1, 10, 15, 21]. Their diffusion models, assuming that the pair-wise influence probabilities are unchanged as the diffusion proceeds, can not capture the characteristics of influence diffusion in Twitter. Their solutions cannot be adopted to the Twitter environment for influence prediction. An early work of analysing patterns of cascading behavior in blog data [13] reports that the most popular shapes of reposting cascades are wide and shallow trees, and the cascade sizes are mostly small. There are similar observations on URL cascades [2, 7] and retweet cascades [20] in Twitter: the majority of messages (ranging from 73% to 95% across domains) in Twitter do not spread at all. This is because that "retweets" and "likes" cannot get any influence credit according to Twitter's mechanism design. In addition, time plays an important role in the diffusion of influence [17]. To predict influence diffusion in Twitter, the above features need to be properly considered. Our proposed diffusion models incorporate the temporal and structural features of influence cascades in Twitter.

3 INFLUENCE IN TWITTER

In this work, we discuss influence in Twitter according to users' actions. *Taking an action* in Twitter is *posting a message*. A message can be a "retweet", a "like", a "reply" or a "quote" (as explicit interactions with other messages); it can also be an independent "tweet" without any explicit relation to any other messages. As discussed earlier "tweets" can have implicit interactions with other messages via the *follow* mechanism. Our definition of influence in Twitter is:

Definition 1. Given an individual user u who posts a message, if a user v reacts to u's message, we define that u has influenced v. The following five kinds of reactions performed by v to u's message are regarded as influence:

- v retweets u's message
- v likes u's message
- v replies to u's message
- v quotes u's message
- v, who follows u, posts a new tweet that with similar contents to u's message within a certain time period (defined in Section 4.1.2).

The ***influence probability*** $p(u, v)$ is defined as the probability that v will react to u's message. $p(u, v)$ is learned from the action history, based on the ratio of the number of u's actions that have influenced v to the total number of u's actions. Based on the pair-wise influence probabilities, we develop influence diffusion models to predict influence cascades [13] in the network, i.e., whom would be influenced by a particular person and when people would be influenced.

4 INFLUENCE PROBABILITY AND DIFFUSION

4.1 Learning Influence Probabilities

We construct the influence network $G_{Infl} = (V, E, P)$, where V is the set of users, E is the set of directed edges, and P is the set of influence probabilities between users, from both the follower network and users' action history. The action history, denoted as $\mathcal{A}$, is a set of all the messages posted by the users in V in a period. $\mathcal{A}_u$ denotes the set of all the messages of a particular user u, including u's "tweets", "retweets", "likes", "replies" , and "quotes", which are denoted as $\mathcal{A}_{u_TWT}$, $\mathcal{A}_{u_RET}$, $\mathcal{A}_{u_LIK}$, $\mathcal{A}_{u_REP}$, and $\mathcal{A}_{u_QUO}$ respectively. $\mathcal{A}_u = \mathcal{A}_{u_TWT} \bigcup \mathcal{A}_{u_RET} \bigcup \mathcal{A}_{u_LIK} \bigcup \mathcal{A}_{u_REP} \bigcup \mathcal{A}_{u_QUO}$. Also, let A_u denote a single message posted by u in $\mathcal{A}_u$.

4.1.1 The influence network structure. For any pair of $(u, v) \in V$, if there is a "follow" relation, or an explicit interaction from v to u, we add a directed edge (u, v) to G_{Infl} to indicate that u potentially influences v. For example, in Figure 1, user H follows B, and thus B has the potential to influence H, and in the derived G_{Infl} (the graph on the right), there is a directed edge (B, H); and H likes G's tweet, so a directed edge (G, H) is added to G_{Infl}.

4.1.2 Learning the implicit influence. As we discussed earlier, people can be influenced implicitly via the *follow* mechanism. The influence can be inferred from the content relations of messages

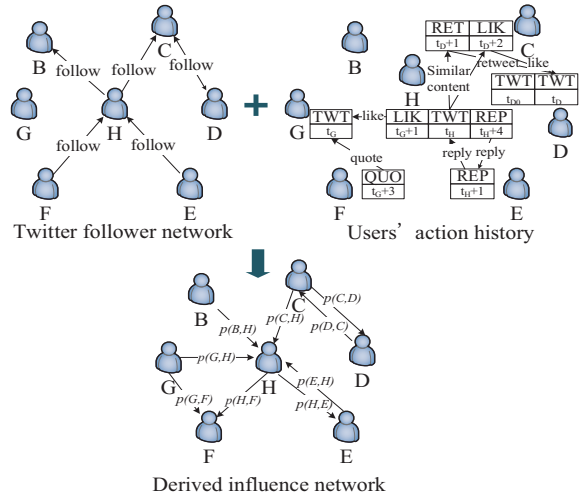


Figure 1: Constructing the influence network

posted by users. The challenge is to identify those tweets that are the results of following the people who post related messages. Tweets in $\mathcal{A}_{v_TWT}$ are classified into two types: the tweets posted by user v without relations to messages of other users, denoted as $\mathcal{A}_{v_TWT_{ini}}$; and the tweets as the results of the influence of the people who v follows, denoted as $\mathcal{A}_{v_TWT_{inf}}$. Our approach of identifying an A_{v_TWT} as an $A_{v_TWT_{inf}}$ considers the social network structure, time and content similarity.

Because of the speed of Twitter streams, messages become quickly dismissed. A concept of *influential time period* T_0 of a message is introduced. We assume that it is the period during which a message can have influence; after that it has no influence. We consider u to have an implicit influence on v, if v, who follows u, posts a new tweet A_{v_TWT} based on u's message A_u within T_0 after u posted the message. After getting all the possible $\langle A_u, A_{v_TWT}\rangle$ pairs for each A_{v_TWT} under the *follow* and T_0 constraints, we use the content similarity to identify whether A_{v_TWT} is an $A_{v_TWT_{inf}}$ based on A_u. To decide whether two messages have the similar content, we check if they have same tokens as user mentions (`@username`) [4], hashtags(`#hashtag`) [18] and urls(`bit.ly`) [2]. If two messages have at least one common token, we regard that their topics are similar and A_{v_TWT} is an $A_{v_TWT_{inf}}$ due to the influence of A_u. If not, we look into the details of these two messages. We remove special tokens, stop words and words with less than 3 letters, and perform stemming and lemmatization on the words. After the pre-processing, the longest message has 34 important words for our data and on average a message contains 8.5 important words (the Standard Deviation is 4.9). Since a message in Twitter is normally short, sharing same important words is an indication that the two messages are about a similar topic. In the *Darwin* dataset collected based on users' location, 92.8% of the $\langle A_u, A_{v_TWT}\rangle$ pairs evaluated by important words have no common words, and 6% have only one common word. We check the message pairs sharing more than two important words and identify whether their topics are similar. It is found that when two Twitter messages share three or more important words, their content are similar with a probability of above 90%. We consider A_{v_TWT} to be an $A_{v_TWT_{inf}}$ if there exists an A_u that shares three or more important words in common with A_{v_TWT}. For an A_{v_TWT}, if there is no A_u satisfying the above three constraints, it is identified as an $A_{v_TWT_{ini}}$.

For an $A_{v_TWT_{inf}}$, there may have more than one A_u. We cannot really know which A_u has influenced v. Facing a similar case in tracking the diffusion of the urls in Twitter, Bakshy et al. [2] propose three choices: "first influence" which assigns full credit to the user who takes action first, "last influence" which assigns full credit to the user who takes action most recently, and "split influence" that splits credit equally among all possible A_u. They report that the qualitative findings are identical though the values varied slightly across the three assumptions. Here we adopt "last influence" and assign full credit to the most recent A_u.

4.1.3 Learning the influence probabilities. We first consider the total number of u's actions $|\mathcal{A}_u|$. Each A_u can influence others implicitly. However, we observe that not every A_u is recorded to have explicit influence. For example, in Figure 1, user G posts a "tweet" at time t_G, H likes it at t_G+1, and F sees the "tweet" from H's "like" and quotes it at t_G+3. F's action is recorded as a reaction to G's "tweet", not H's "like" of G's "tweet". The reactions to a "retweet" or "like" will be counted to the user who posts the root message. Only "tweet", "quote" and "reply" are recorded to have explicit influence on others. We refer to them as "Explicit Influenceable Actions"(EIA), denoted as $\mathcal{A}^*$: $\mathcal{A}^*_u = \mathcal{A}_{u_TWT} \bigcup \mathcal{A}_{u_REP} \bigcup \mathcal{A}_{u_QUO}$. Both the number of total actions $|\mathcal{A}_u|$ and EIAs $|\mathcal{A}^*_u|$ are learned when scan the action history.

Then we consider the number of u's actions that have influenced v. Based on $\mathcal{A}^*_u$ and $|\mathcal{A}_u|$, we learn u's actions that have influenced v in explicit ways, denoted as $\mathcal{A}_{u2v_e}$, and u's successful actions influencing v in implicit ways, denoted as $\mathcal{A}_{u2v_i}$, separately in the scanning. In Twitter, v is able to react directly with u's same message multiple times in various ways. For example, in Figure 1, user C has two reactions (liking and retweeting) to D, and both of these two reactions are to D's "tweet" posted at t_D. Only one of D's actions has influenced C. $|\mathcal{A}_{u2v_e}|$ or $|\mathcal{A}_{u2v_i}|$ is the number of u's unique actions that have influenced v explicitly or implicitly. It requires just one scan of the action history to get $|\mathcal{A}^*_u|$, $|\mathcal{A}_u|$, $|\mathcal{A}_{u2v_e}|$ and $|\mathcal{A}_{u2v_i}|$.

We calculate $p(u,v)$ based on the ratio of the number of u's actions that have influenced v to the total number of u's actions specified in Section 3. A user u's explicit influence on v, denoted as $p^e_{u,v}$, and implicit influence on v, denoted as $p^i_{u,v}$, are calculated separately as:

$$p^e_{u,v} = f(|\mathcal{A}^*_u|) \cdot \frac{|\mathcal{A}_{u2v_e}|}{|\mathcal{A}^*_u|} \tag{1}$$

$$p^i_{u,v} = f(|\mathcal{A}_u|) \cdot \frac{\beta|\mathcal{A}_{u2v_i}|}{|\mathcal{A}_u|} \quad \beta \in (0,1) \tag{2}$$

In Equation 1 and 2, if the denominator $|\mathcal{A}^*_u|$ or $|\mathcal{A}_u|$ is small, the ratio can be high (e.g. $\frac{|\mathcal{A}_{u2v_e}|}{|\mathcal{A}^*_u|} = \frac{1}{1} = 1$). A small denominator means insufficient observations of actions, and we employ $f(x) = 1/(1+e^{(-(x-1))})$ as a scaling factor (see Figure 2) to adjust its effect. $f(x)$ is 0.5, 0.73, and 0.88 when x is 1, 2 and 3 respectively; $f(x)$ is 0.95 when x 4 and approaches to 1 while x is bigger. In Equation 2, a control parameter β in the range $(0,1)$ is applied on the implicit

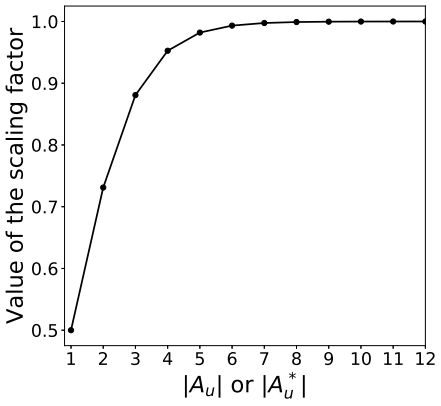


Figure 2: The scaling factor in Eq. 1 and 2

influence. We calculate p_2 with the joint probability of $p^e_{u,v}$ and $p^i_{u,v}$:

$$p_2 = 1 - (1 - p^e_{u,v})(1 - p^i_{u,v}) \tag{3}$$

It is assumed that there is always a small influence probability between connected user pairs in the influence network, which is denoted as p_1. $p(u, v)$ is calculated as:

$$p(u, v) = 1 - (1 - p_1)(1 - p_2) \tag{4}$$

$p(u, v)$ is time-related if T_0 constraint is applied to get $|\mathcal{A}_{u2v_e}|$ and $|\mathcal{A}_{u2v_i}|$. $p^{T_0}(u, v)$ is the probability that u's action will influence v in T_0's time since u takes the action; after T_0, the probability drops to 0. To set $|\mathcal{A}_{u2v_i}|$ as 0, we get the probability that u influences v without considering the implicit influence.

4.2 Diffusion Models

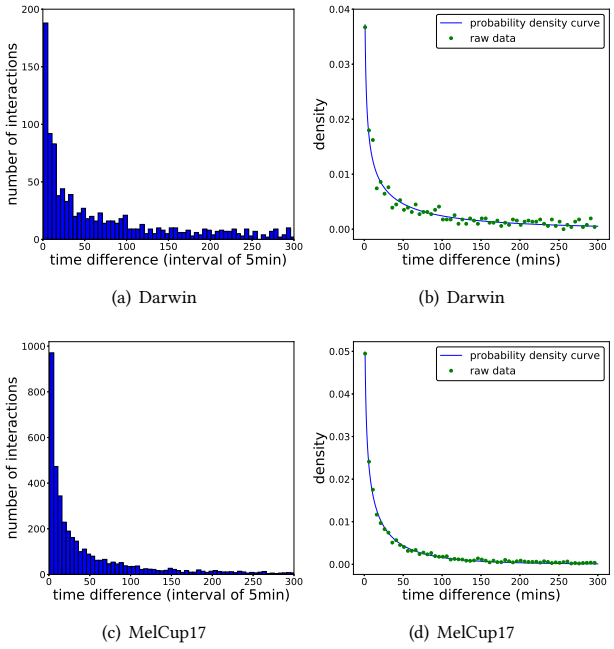
We apply the pair-wise influence probabilities to diffusion models to predict how influence diffuses among Twitter users. In the real world, the diffusion of influence is continuous, while cascade model proceeds in discrete steps from one time slot to the next. We assume each time slot is equal and choose the time slot t_s based on the temporal features of influence in learning data. After u takes an action, u is able to influence his/her out-degree neighbour v who has not been influenced regarding to this action in a maximum of $N = |\frac{T_0}{t_s}|$ steps, with a total probability of $p^{T_0}(u, v)$ and a probability $p_i(u, v)$ at the i^{th} step after u takes the action. We have

$$p^{T_0}(u, v) = 1 - \prod_{i}^{N}(1 - p_i(u, v)) \qquad (i = 1, 2, ..., N) \tag{5}$$

After N steps (i.e. a time period of T_0), u's influence on its out-degree neighbours drops to 0: $p_i(u, v) = 0 \quad (i > N)$. For an action, its influence cascade terminates if no more user is influenced.

For v who has not been influenced, denote S as the set of v's in-degree neighbours who are able to influence v. If there are multiple $u_j \in S$, we assume that the influence from multiple neighbours to v are independent of each other and adopt a joint probability $p_v(S)$, proposed by [9], to define the probability that v will be successfully influenced:

$$p_v(S) = 1 - \prod_{u_j \in S}(1 - p_i(u_j, v)) \tag{6}$$



(a) Darwin (b) Darwin

(c) MelCup17 (d) MelCup17

Figure 3: (a/b). Frequency of interactions vs. the time difference between interactions of *Darwin/MelCup17* (c). Fitting the frequency of interactions over time of *Darwin*

4.2.1 Time Constant Cascade Model(TC-C). In the Time Constant Cascade model, $p_i(u, v)$ is independent of time, which means at each time step i ($i \leq N$) after u takes the action, $p_i(u, v)$ stays the same.

$$p_i(u, v)_c = p_{i+1}(u, v)_c \qquad (i = 1, 2, ..., N - 1) \tag{7}$$

Then the constant $p_i(u, v)_c$ is solved from (5) and (7) as:

$$p_i(u, v)_c = 1 - (1 - p^{T_0}(u, v))^{1/N} \tag{8}$$

After i steps, the total probability that u influencing v is:

$$p^i(u, v)_c = 1 - \prod_{1}^{i}(1 - p_i(u, v)_c) \qquad (i = 1, 2, ..., N) \tag{9}$$

Because $p_i(u, v)_c = p_{i+1}(u, v)_c$, the joint influence probability from multiple neighbours when a new influential neighbour w adding in $p_v(S \bigcup w)$ and when a influential neighbour w becomes having no influence $p_v(S \setminus w)$ in TC-C models could be updated incrementally from $p_v(S)$ [9].

4.2.2 Time Decay Cascade Model(TD-C). We plot the number of explicit interactions against the time to show when v is influenced by u. The statistics of *Darwin* and *MelCup17* datasets are given in Figure 3(a) and 3(c) respectively. Each figure shows the frequency of interactions in intervals of 5 minutes over a span of 300 minutes. The majority of the interactions occur within a very short period of time and the number of interactions decay quickly. User interactions

in Twitter are generally fast and short in lifespan [14]. We assume that for each (u, v) pair in a social network, $p(u, v)$ decays over time following the same fashion. To identify how $p(u, v)$ decays, we fit a probability density curve $p(x)$ based on the interaction frequency distribution over time, see Figure 3(b) and 3(d).

In the TD-C model, the relationships among $p_i(u, v)_d$ are as follows:

$$\frac{p_1(u,v)_d}{\int_0^{t_s} p(x)\,dx} = \ldots = \frac{p_i(u,v)_d}{\int_{(i-1)\cdot t_s}^{i\cdot t_s} p(x)\,dx} \qquad (i = 1, 2, \ldots, N) \quad (10)$$

After i steps, the total probability that u influencing v is:

$$p^i(u,v)_d = 1 - \prod_1^i (1 - p_i(u,v)_d) \qquad (i = 1, 2, \ldots, N) \quad (11)$$

With (5) and (10), the values of $p_i(u, v)_d$ can be obtained. Since $p_i(u, v)_d$ changes as the diffusion proceeds, the $p_v(S \bigcup w)$ and $p_v(S \setminus w)$ can not be computed incrementally.

For any $i \in [1, N-1]$, $p^i(u, v)_d$ in (11) is larger than $p^i(u, v)_c$ in (9). Due to limited space, we omit the proof here. It results that for the same $p^{T_0}(u, v)$, at any step before no more users can be influenced, the number of influenced users with the TD-C model will always be larger than that with the TC-C model.

4.2.3 Time-Depth Decay Cascade Model(TDD-C). Existing studies [7, 8] report that over 99% of the interactions to the root user in Twitter take place within one depth of the root user and the depth distribution of diffusion cascade graph is skewed. We consider the decay of influence with the influence diffusion cascade depth. Suppose that the shortest path length of an influenced user v to the root user u is $l(u, v)$, then in the TDD-C model the influence probability of v to its neighbour w, who has not been influenced at the i^{th} step since v takes the action, is calculated as

$$p_i(v,w)_{dd} = \delta^{l(u,v)} p_i(v,w)_d \quad (12)$$

δ is a decay factor for depth, ranging from 0 to 1. Only "reply" and "quote" to a root action can make the cascades go further and δ is learned as the ratio of the sum of "reply" and "quote" to the total number of explicit interactions.

4.2.4 Model Variations. With each of the three models, there are two variations depending on how $p(u, v)$ is learned. In total, there are six models: TC-Cp, TC-Cp_e, TD-Cp, TD-Cp_e, TDD-Cp, and TDD-Cp_e. p stands for the proposed models with probabilities learned from both explicit and implicit interactions, and p_e stands for the ones with probabilities learned from the explicit interactions only.

5 EVALUATION

TC-Cp_e has been studied in [9, 15] and is used as the baseline. Starting from an action initiated (i.e. a "tweet") by a root user u, the proposed diffusion models are used to simulate how this tweet is propagated through users in the influence network. With multiple rounds of simulations, the results give answers to the following questions: (a). On average, how many users can be influenced by u? (b). Who are the users that will be influenced by u? (c). When are these users influenced? (d). What are the shapes of the influence diffusion cascades like?

1. Predicting size The size of an influence diffusion cascade is the number of users involved in the cascade [13]. For a user u, $s(u)$ is the average number of users in the influence cascades of actions initiated by u according to the historical data, and $\hat{s}(u)$ is the predicted value. The difference between $s(u)$ and $\hat{s}(u)$ is represented with Normalise Mean Squared Error (NMSE):

$$NMSE = \frac{1}{n}\sum_i \left(\frac{s(u_i) - \hat{s}(u_i)}{s(u_i)}\right)^2$$

2. Predicting Shape For the influence cascade shape, we evaluate the depth distribution of the influenced users in the influence cascades graph. The depth of an influenced user is the shortest path length between this user and the root.

3. Predicting time The actual time v is influenced by an action is t_v and the $\hat{t}_v$ is the time that v is predicted to be influenced. In the applications of influence diffusion, such as viral marketing, the focus is how many users are influenced within a certain period of time [9]. If $\hat{t}_v \le t_v$, we consider that the model successfully predicts when v will be influenced.

4. Predicting who influences whom at what time For each A_{u_TWT} in the evaluation data, its actual influence is built up in the form of $(A_{u_TWT}, u, v, v_{infled})$. v_{infled} marks whether v is influenced by A_{u_TWT}: it is 0 if v is not influenced, and if v is influenced, its value is the influenced time t_v. All users that have been influenced by A_{u_TWT} are denoted as $U_1(A_{u_TWT})$. The out-degree neighbours of $U_1(A_{u_TWT})$ and u in the influence network who are not influenced by A_{u_TWT} are the users whose $v_{infled} = 0$), denoted as $U_0(A_{u_TWT})$.

The simulated results of a simulated diffusion starting from an A_{u_TWT} are presented as a form of $(A_{u_TWT}, u, \hat{v}, \hat{v}_{infled} = \hat{t}_v)$. Note that the diffusion models proceed in discrete steps from one time slot t_s to the next. A user is influenced at i^{th} ($i >= 1$) time slot, i.e. during a time period of $[(i-1)\cdot t_s, i\cdot t_s]$. We let $\hat{t}_v = (i-1)\cdot t_s$. The influenced users in the simulations are denoted as $\hat{U}_1(A_{u_TWT})$.

For each action A_{u_TWT}, we get the TP, FN, FP, TN cases of predicting whom are influenced or not as follows: for each $v \in U_1(A_{u_TWT}) \bigcup U_0(A_{u_TWT})$,

- if $v_{infled} > 0$ and $v \in \hat{U}_1(A_{u_TWT})$, it is a TP;
- if $v_{infled} > 0$ and $v \notin \hat{U}_1(A_{u_TWT})$, it is a FN;
- if $v_{infled} == 0$ and $v \in \hat{U}_1(A_{u_TWT})$, it is a FP;
- if $v_{infled} == 0$ and $v \notin \hat{U}_1(A_{u_TWT})$, it is a TN.

By considering if a model predicts when v is influenced, the criteria of TP cases are more rigorous:

- if $v_{infled} > 0$ and $v \in \hat{U}_1(A_{u_TWT})$ and $\hat{t}_v \le t_v$, it is a TP;
- if $v_{infled} > 0$ and $v \in \hat{U}_1(A_{u_TWT})$ and $\hat{t}_v > t_v$, it is a FN.

In Twitter, $|U_1(A_{u_TWT})|$ is relatively small compared to $|U_0(A_{u_TWT})|$. As a result, the number of positive cases is much smaller than the number of negative cases in the actual data (TP, FN and FP cases will be much smaller than TN). The *precision*, *recall* and *F-score* measures are used to evaluate the prediction performance.

6 EXPERIMENTS

6.1 Datasets

This work learns the influence probabilities between users from their historical action data. Existing open datasets [6, 22] are not suitable for this purpose due to that they have not included rich

Table 2: Description of the Actions and Interactions

Datasets	action data				interaction data	
	#statuses	#likes	#actions	#users	#intractions	#users
Darwin-lrn	69,323	60,481	129,804	1,265	6,793	437
Darwin-eva	47,749	40,550	88,229	1,130	3,127	375
MelCup17-lrn	12,977	11,024	24,001	1,484	12,894	1,391
MelCup17-eva	10,567	9,266	19,833	1,374	7,670	1,132

Table 3: Description of the Influence Network

Dataset	following network		interaction network		influence network	
	# users	# edges	# users	# edges	# users	# edges
Darwin	3,880	58,503	437	2,080	3,883	58,799
MelCup17	1,484	70,756	1,391	7,403	1,484	72,710

interactions. We collect two datasets [1] from Twitter: one is based on users' location, and the other one is based on an event.

Location-based Dataset-"*Darwin*" We first identify 6,571 Twitter users whose location are "Darwin, Northern Territory, Australia" and collect their data real time from 16/11/2017 to 27/12/2017, Australian Central Standard Time(ACST) for learning (*Darwin-lrn*). The data includes "tweet", "retweet", "reply", and "quote", which are named as "statuses". There are 69,323 statuses, including 19,164 tweets. We then collect 60,481 "likes" of these users in this period. In total, we get 129,804 messages/actions posted/taken by 1,265 users. There are 6,793 (excluding self-interactions) interactions among 437 users, resulting an interaction network of 2,080 edges, see details in Table 2. We further collect a dataset of the same users from 30/12/2017 to 26/01/2018, ACST, as the ground truth for evaluation (*Darwin-eva*). It contains 47,749 statuses and 40,550 likes of 1,130 users. We then collect the *following* relations among those users. There are 58,503 relations among 3,880 users. From the interactions and following networks, we get the influence network. It includes 3,883 users and 58,799 edges (see Table 3).

In the *Darwin-eva* dataset, there are 13,656 tweets initiated by 858 users; 660 users who initiate 9,684 actions influence no one in the *Darwin* community; 198 users have 833 tweets out of their total 3,972 tweets that successfully influenced others. For these 833 tweets, on average, there are 2.80 users and 3.18 actions in an influence cascade, and the largest cascade contains 16 users and 20 actions. The average cascade size of users is smaller than the size of actions because each user can participate multiple times in a cascade. We focus on whom would be influenced, and the cascade size reported in the results is the number of users. More descriptions of the evaluation data will be provided in the results part.

Event-based Dataset-*"MelCup17"* We use Twitter *Streaming API* to track the terms "Melbourne Cup", "#MelCup" and "Derby-Day", from 00:00:00, 06/11/2017 to 23:59:59, 08/11/2017, Australia Eastern Standard Time (AEST), covering three days around the 2017 Melbourne Cup Day. 78,300 statuses messages posted by 34,573 users during the three days are collected. The results of Melbourne Cup came out on 15:05:19, 07/11/2017, AEST. The data before the results is used in the learning (*MelCup17-lrn*), and the data produced after the results is used for the evaluation (*MelCup17-eva*). We filter the users who have less than 6 statuses. As a result, 23,544 statuses posted by 1,484 users are investigated. The "likes" and *following* relations of the 1,484 users during this period are then collected. Descriptions of the *"MelCup17"* dataset are given in Table 2 and 3.

In the *MelCup17-eva* dataset, there are 3,194 tweets initiated by 740 users. 388 users who initiate 1,713 tweets influence no one in the *MelCup17* community. And 352 users have 1,052 tweets, out of their total 1,481 tweets, that successfully influence others. For the 1,052 tweets influencing others, the average cascade size of users is 5.5 and the largest cascade contains 89 users.

[1] The datasets and codes are available at https://bit.ly/2UlQ3xW

Table 4: Predicting size

Dataset	parameters	TC-Cp	TD-Cp	TDD-Cp
Darwin	α=1.2,t_s=15	1.1594	4.7386	0.7550
	α=1,t_s=15	**0.2227**	0.4954	**0.1665**
	α=0.8,t_s=15	0.2368	**0.2456**	0.1758
MelCup17	α=1,t_s=10	1.1869	9.1994	**0.4876**
	α=0.9,t_s=10	**0.6593**	1.6468	0.5030
	α=0.7,t_s=10	0.7012	**0.7186**	0.6343

6.2 Experimental Setup

In our experiments, T_0 is set as 300 minutes based on the observation that 80% of the reactions to the root action happen in 300 minutes.β is set as 0.5 and we choose p_1 as 0.001. The depth decay factor δ is learned as 0.15 and 0.13 for the "*Darwin*" and "*MelCup17*" respectively. Different t_s ranging from 3mins to 15mins are used. For different diffusion models, different scales of the learned $p^{T_0}(u,v)$s are used to adjust the sizes of the simulated influence cascades to fit the actual cascade sizes. The values of the scaling factor α are learned for different models. Under a set of parameters, we simulate the influence diffusion starting from each user who has at least an A_{u_TWT} in the evaluation data, with each model 100 rounds.

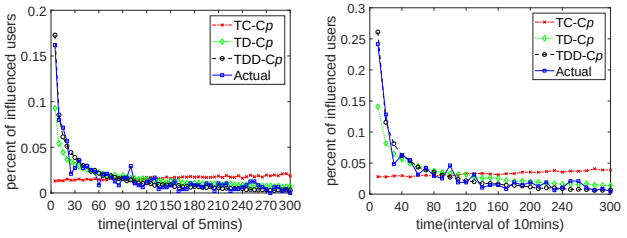
6.3 Experimental Results

Size Our analysis starts from the simulated cascade size $\hat{s}(u)$. The number of users a user influences when initiating an action is the most straightforward measure of how well the predictions fit what actually happened. We first figure out a proper α to make $\hat{s}(u)$ close to $s(u)$ in different models. Table 4 shows the NMSE values between $\hat{s}(u)$ and $s(u)$ under different α in TC-Cp, TD-Cp, and TDD-Cp models. With a same α, $\hat{s}(u)$ in the TD-C model is always the largest (see Section 4.2.2). For the *Darwin* dataset, when $\alpha = 1.2$, the $\hat{s}(u)$ values are generally larger than the $s(u)$, and the NMSEs are large. With $\alpha = 1$, the TC-Cp and TDD-C$_p$ predict the number of users a user will influence with low errors. And the TD-Cp model has similar performance when $\alpha = 0.8$. For the *MelCup17* dataset, the α is 0.9, 1 and 0.7 in the TC-C, TD-C, and TDD-C models respectively.

Shape Table 5 shows the statistics of the depth distribution of the influenced users with the three models ($t_s = 5mins$ for *Darwin*, $t_s = 3mins$ for *MelCup*17). The results for different t_s are similar. The prediction results with TC-C or TD-C models have a larger fraction of users influenced more than 1-depth away from the root user. The results with TDD-C model fit the depth distribution of influence cascades in the real world better than other models.

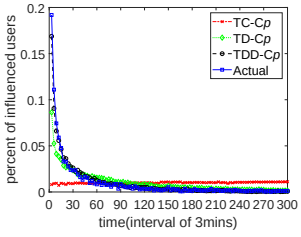
Table 5: Predicting shape

Dataset	Depth	Actual	TC-Cp	TD-Cp	TDD-Cp
Darwin	1	95.9%	86.82%	73.16%	94.64%
	2	2.4%	13.10%	26.05%	5.35%
	3	1.4%	0.08%	0.79%	0.01%
MelCup17	1	98.2%	68.67%	64.66%	95.13%
	2	1.8%	30.36%	33.96%	4.87%
	3	0.0%	0.97%	1.38%	0.00%

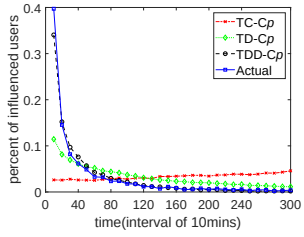


(a) *Darwin*, t_s=5mins

(b) *Darwin*, t_s=10mins



(c) *MelCup17*, t_s=3mins



(d) *MelCup17*, t_s=10mins

Figure 4: The time when users are influenced

Whom would be influenced by a particular person Table 6 shows the precision, recall and F-score values achieved with different models. For each A_{u_TWT}, its actual influence cascade is compared with 100 simulated cascades with each model, and the result with the best F-score is adopted. The values are the average of the results of all the A_{u_TWT}. Results with t_s of 5, 10, and 15 minutes in the *Darwin* dataset are given. The best performances achieved by the models are mainly with t_s=5mins, and the variations with different t_s are small. For brevity, we only provide the values with t_s=3min for the *MelCup17* dataset.

In general, the proposed models with p learned from both explicit and implicit interactions perform better than the models with p_e learned only from the explicit interactions. We provide a case study in *Darwin*: user $u2273$ and his follower $u2114$. In the learning data, $u2114$ has no explicit reactions to $u2273$, but 22 of $u2273$'s actions influence $u2114$ implicitly. In the evaluation data, $u2114$ is observed liking or retweeting eight of $u2273$'s messages. Without considering implicit influence, the models with very small $p_e(u2273, u2114) = 0.001$ can not predict that $u2114$ would be influenced by $u2273$ in most cases. With $p(u2273, u2114) = 0.152$, the models predict that $u2114$ is influenced by $u2273$ 20 times averagely in the 100 rounds of simulations. Our algorithm of learning implicit influence in Twitter based on *follow* relations, time and content similarity is effective.

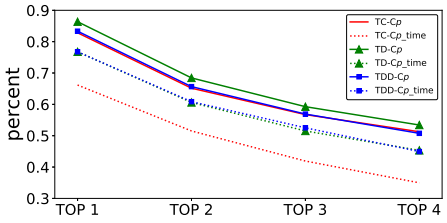


Figure 5: Top N predicted being influenced users in actual data

All the three types of models work in predicting whom would be influenced, referring to the left rows of the models in Table 6. In the *Darwin* data, when $t_s = 5mins$, the TDD-Cp model achieves the best precision of 0.7016. The highest recall goes to the TD-Cp model, which is 0.6584. The TDD-Cp model gives the best F-score of 0.6299. With different t_s, the TDD-Cp model always gives the best performance in precision and F-score measures, which shows that our proposed time and depth-sensitive diffusion model works. All our five models outperforms the baseline TC-Cp_e model.

When people would be influenced Figure 4 shows at what time the users are influenced in the evaluation data and the proposed models with different t_s. It shows the percentage of influenced users against the time that elapsed between t_v (the time when v is influenced) and t_u (the time when u initiate an action). The curves of each model's two variations are quite similar, and we only show the ones of models with p. In the actual situation, the number of influenced users decays over time. Both the TD-C and TDD-C models catch the decay behaviour. In terms of the percentage, the results of TDD-C fit the actual data better. The flat curve of TC-C model demonstrates that its prediction is not accurate.

The values of the prediction metrics calculated by considering when people are influenced are given in the right columns of different models in Table 6. To keep consistent in different t_s, the time related metrics are all calculated in a granularity of 30 minutes. We can see that the values of the prediction metrics of TC-C models drop significantly from above 0.6 to around 0.15. While the values with TD-C and TDD-C models also drop, to 0.40-0.50, they still show a good performance. In the *Darwin* dataset, when t_s=5mins, the precision, recall, F-score of the TDD-Cp model outperform the baseline TC-Cp_e model 219%, 244%, and 234% respectively. For the *MelCup17*, the numbers are 203%, 153%, and 186% when t_s=3mins. Our approach of modelling the influence diffusion in Twitter with time-sensitive probabilities is effective.

We are interested in the users who are influenced most times by a particular person. Figure 5 presents the percentage of the Top 1-4 predicted being influenced users that are actually influenced in the *Darwin* evaluation dataset with $t_s = 5min$. Looking at who are the most influenced user, the TD-Cp model performs slightly better than TC-Cp and TDD-Cp model. On average, 84.2% of the Top 1 predicted influenced users with the three models are actually influenced ; and the percentages are 66.4%, 57.6% and 51.8% for the Top 2, 3 and 4. Considering when the top users are influenced, the performance of TC-Cp model drops by 25%. While the performances of TD-Cp and TDD-Cp model drop by around 10%.

Table 6: Performance of predicting who will influence whom at what time

Data& metrics		TC-Cp		TC-Cpe (Baseline)		TD-Cp		TD-Cpe		TDD-Cp		TDD-Cpe	
		who	when	who	when	who	when	who	when	who	when	who	when
Darwin 5min	Precision	0.6618	0.1810	0.6581	0.1517	0.6593	0.4168	0.6454	0.4077	**0.7016**	***0.4845***	0.6910	0.4791
	Recall	0.6178	0.1627	0.6017	0.1251	0.6405	0.4023	**0.6584**	0.3912	0.6300	***0.4300***	0.6074	0.4122
	F-score	0.6079	0.1621	0.5978	0.1290	0.6107	0.3843	0.6082	0.3746	**0.6299**	***0.4303***	0.6159	0.4193
Darwin 10min	Precision	0.6759	0.1720	0.6635	0.1935	0.6637	0.4294	0.6368	0.4179	**0.7033**	***0.4750***	0.6888	0.4577
	Recall	0.6279	0.1506	0.6223	0.1628	**0.6647**	***0.4187***	0.6421	0.4008	0.6125	0.4147	0.6147	0.3978
	F-score	0.6171	0.1509	0.6094	0.1672	0.6208	0.3971	0.5990	0.3832	**0.6231**	***0.4185***	0.6187	0.4031
Darwin 15min	Precision	0.6809	0.1927	0.6579	0.2137	0.6569	0.4115	0.6522	0.4189	**0.6999**	***0.5539***	0.6773	0.5318
	Recall	0.6424	0.1707	0.6151	0.1935	**0.6666**	0.4066	0.6432	0.3962	0.6161	***0.4766***	0.6005	0.4611
	F-score	**0.6273**	0.1714	0.6060	0.1937	0.6150	0.3828	0.6061	0.3818	**0.6249**	***0.4884***	0.6058	0.4688
MelCup17 3min	Precision	0.5707	0.1539	0.5685	0.1445	0.5660	0.3840	0.5320	0.3365	**0.6112**	***0.4374***	0.6055	0.3891
	Recall	0.5485	0.1459	0.5310	0.1526	0.5520	0.3765	0.5393	0.3595	**0.5641**	***0.3861***	0.5133	0.3152
	F-score	0.4918	0.1282	0.4906	0.1263	0.4921	0.3308	0.4690	0.2976	**0.5223**	***0.3615***	0.4940	0.3032

7 CONCLUSION

In this paper we addressed the questions of measuring people's influence on others and predicting influence diffusion in Twitter. We proposed an effective algorithm for learning influence in Twitter. By capturing both the explicit and implicit influence in the action history, the pair-wise influence probabilities have been learned. We then proposed a time and depth-sensitive cascade model for influence diffusion by considering the temporal and structural features of actual influence cascades in Twitter. Experiments are conducted on two real-world Twitter datasets. Results show that our proposed models can predict the influence cascade in Twitter quite well, outperforming the existing models to predict when the users are influenced in terms of the balanced precision and recall. Our prediction task is important for understanding how influence diffusion works with many applications, such as marketing of time-sensitive goods and recommending the important messages that users are most likely to adopt in the information overloaded social networks.

Recent work has investigated topic-wise social influence diffusion [3]. It would be interesting to look at user-to-user influence and the spreading of influence in specific topics or domains.

REFERENCES

[1] Jeffery Ansah, Wei Kang, Lin Liu, Jixue Liu, and Jiuyong Li. 2018. Information Propagation Trees for Protest Event Prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 777–789.

[2] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 65–74.

[3] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2013. Topic-aware social influence propagation models. *Knowledge and information systems* 37, 3 (2013), 555–584.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. 2010. Measuring user influence in twitter: The million follower fallacy. *Icwsm* 10, 10-17 (2010), 30.

[5] Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. 2013. Information and influence propagation in social networks. *Synthesis Lectures on Data Management* 5, 4 (2013), 1–177.

[6] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. 2013. The anatomy of a scientific rumor. *Scientific reports* 3 (2013), 2980.

[7] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.

[8] Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. 2012. The Structure of Online Diffusion Networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, New York, NY, USA, 623–638.

[9] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 241–250.

[10] Siwar Jendoubi, Arnaud Martin, Ludovic Liétard, Hend Ben Hadji, and Boutheina Ben Yaghlane. 2017. Two evidential data based models for influence maximization in twitter. *Knowledge-Based Systems* 121 (2017), 58–70.

[11] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[12] Konstantin Kutzkov, Albert Bifet, Francesco Bonchi, and Aristides Gionis. 2013. Strip: stream learning of influence probabilities. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 275–283.

[13] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2007. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 551–556.

[14] Bjoern-Elmar Macek and Martin Atzmueller. 2013. Visualizing the impact of time series data for predicting user interactions. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 1477–1478.

[15] Yan Mei, Weiliang Zhao, and Jian Yang. 2017. Influence maximization on twitter: A mechanism for effective marketing campaign. In *Communications (ICC), 2017 IEEE International Conference on*. IEEE, 1–6.

[16] Sancheng Peng, Yongmei Zhou, Lihong Cao, Shui Yu, Jianwei Niu, and Weijia Jia. 2018. Influence analysis in social networks: A survey. *Journal of Network and Computer Applications* (2018).

[17] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1 (2014), 26–65.

[18] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 695–704.

[19] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. 2008. Prediction of information diffusion probabilities for independent cascade model. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 67–75.

[20] Io Taxidou and Peter M. Fischer. 2014. Online Analysis of Information Diffusion in Twitter. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. ACM, New York, NY, USA, 1313–1318.

[21] Yu-Ting Wen, Wen-Chih Peng, and Hong-Han Shuai. 2018. Maximizing Social Influence on Target Users. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 701–712.

[22] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1513–1522.

Workshop on Subjectivity, Ambiguity and Disagreement on the Web (SAD2019) Chairs' Welcome

It is our SAD pleasure to welcome you to the *WWW 2019 Workshop on Subjectivity, Ambiguity and Disagreement on the Web (SAD2019)*. The first SAD workshop was held in 2018 in Zurich, as part of the HCOMP conference, and had more than 30 attendees, lively discussion and excellent keynotes by Brent Hecht (Northwestern University) and Drazen Prelec (MIT). TheWebConf 2019 has allowed us to draw in a more diverse audience of people from IR, Machine Learning, Recommender Systems, Web Science, etc., who have experience with the issues of disagreement in human-collected data, as well as our Crowdsourcing base. A new goal for this second workshop is to begin to *collect data* and *establish challenges* for productively using data sets that exhibit disagreement.

Ambiguity creates uncertainty in practically every facet of the web. Machine learning, recommender systems, web search, as well as news and systems that support human discourse, experience aspects of ambiguity and subjectivity that lead to inaccuracies, poor and undesirable performance. Recent research has attempted to delve deeper into subjectivity and ambiguity, finding it to manifest in human-collected data primarily as disagreement [Cheatham and Hitzler, 2014; Plank, Hovy and Sogaard, 2014; Bayerl and Paul, 2011; Aroyo and Welty, 2015; Schaekermann, Law, Williams and Callaghan, 2016; Chang, Amershi and Kamar, 2017; Lin and Weld, 2014; etc.]. This includes the information presented to workers as part of a crowdsourcing task, the instructions for what to do with it, and the information they are asked to provide. These ambiguities become deeply tied into our machine learning models and metrics, as they are in the gold data. Similar ambiguity is found in interpreting and deriving utility from user generated data from large scale systems such as social media and search engines. Disagreement surfaces in collaborative projects such as Wikipedia, online forums, and semantic markups. In language, ambiguity can result from missing details, contradictions and subjectivity. Subjectivity may stem from differences in cultural context, life experiences, or individual perception of hard-to-quantify properties. All of these can leave people with conflicting interpretations, leading to results that system builders would regard as "wrong".

The workshop papers present theoretical work, algorithms, applications, systems and data from a diverse range of empirical studies touching on topics such as modeling disagreement in user generated data and content, disagreement in human evaluation of large-scale systems, the relation between disagreement, ambiguity and subjectivity, costs and challenges introduced by ambiguity, designing for tasks with high subjectivity and low inter-rater reliability. We touch on semantic, linguistic, commonsense, and moral judgments, better metrics for characterizing disagreement, ambiguity in human computation task design.

This year, for the first time, the SAD workshop ran a *joint keynote series in collaboration with the HumBL2019 workshop*[1] featuring five keynotes from industry and academia. "What we talk about when we talk about crowdsourcing" by Maria Stone (Apple) and "Humans, machines and disagreement: lessons from production" by Brad Klingenberg (StitchFix) were the two keynotes presented at SAD2019. The keynotes by Anima Anandkumar from NVIDIA, Anurag Batra from Google AI and Jon Chamberlain from the University of Essex were presented at HumBL2019.

This year, we accepted six papers from all around the world covering a broad range of topics. We evaluated them regarding their relevance, quality, and novelty with respect to the topics of the workshop call for papers. You can find the full workshop program at the workshop website[2]. Below we summarize the accepted contributions presented at the workshop.

- *Discovering User Bias in Ordinal Voting Systems*, by Lees et al, proposes methods for de-biasing inherently subjective user judgments, such as ratings of restaurants, films, etc by

[1] https://humlworkshop.github.io/HumBL-WWW2019/

[2] https://sadworkshop.wordpress.com

identifying consistent user voting bias and harnessing its signal for more reliable and/or informative ratings.

- *TaskMate: A mechanism to improve the quality of instructions in crowdsourcing*, by Manam et al, discusses different proactive and reactive methods for resolving ambiguities in instructions while allowing crowdsourced workers to collaborate on improving the quality of instructions, offloading the work from the requester.
- *Learning to Predict Population-Level Label Distributions*, by Liu et al, reduces the total number of human labels needed per item by grouping semantically similar items, pooling their labels, and producing a single label distribution that can be shared by all members of the class, tested on two Twitter corpora.
- *Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions*, by Aroyo et al, provides a survey on the current state of the art on detecting negative online behaviors and outlines open questions and challenges, especially of how toxicity can be more reliably assessed and ranked through a better understanding of personal, cultural, and societal perspectives, given that human judgment of toxicity is an inherently subjective task.
- *Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing*, by Jiménez et al, compares intra- and inter-rater agreement of speech quality judgments in lab settings vs. crowdsourced, showing crowdsourcing to be a valid method for collecting reliable assessments of speech quality, and that disagreement can be used as a signal for difficult cases for annotators to evaluate.
- *Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format*, by Schaekermann et al, presents results from a pilot experiment with an adjudication system for medical-rating task of sleep stages, by eliciting expert disagreement and allowing experts to provide their labeling rationale.

Lora Aroyo
Workshop organizer
Google

Mike Schaekermann
Workshop organizer
University of Waterloo

Anca Dumitrache
Workshop organizer
FD Mediagroep

Jennimaria Palomaki
Workshop organizer
Google

Alexander Quinn
Workshop organizer
Purdue University

Michael Tseng
Workshop organizer
Google

Olivia Redfield
Workshop organizer
Jigsaw

Praveen Paritosh
Workshop organizer
Google

Chris Welty
Workshop organizer
Google

Workshop on Subjectivity, Ambiguity and Disagreement on the Web (SAD2019) Organization

Organization Committee: Lora Aroyo *(Google, USA)*
Mike Schaekermann *(University of Waterloo, Canada)*
Anca Dumitrache *(FD Mediagroep, NL)*
Jennimaria Palomaki *(Google, USA)*
Michael Tseng *(Google, USA)*

Alexander Quinn *(Purdue University, USA)*
Olivia Redfield *(Google, USA)*
Praveen Paritosh *(Google, USA)*
Chris Welty *(Google, USA)*

Program Committee:

Peter Organisciak *(University of Denver, USA)*
Maria Stone *(Apple, USA)*
Omar Alonso *(Microsoft, USA)*
Mike Schaekermann *(University of Waterloo, Canada)*
Anca Dumitrache *(FD, NL)*
Jennimaria Palomaki *(Google, USA)*
Michael Tseng *(Google, USA)*
Alexander Quinn *(Purdue University, USA)*
Olivia Redfield *(Google, USA)*
Chris Welty *(Google, USA)*

Humans, Machines and Disagreement

Lessons from Production

Brad Klingenberg[†]
Stitch Fix
San Francisco, California, USA, brad@stitchfix.com

ABSTRACT

Most recommendation algorithms produce results without humans-in-the-loop. Combining algorithms with expert human curation can make recommendations much more effective, especially in hard-to-quantify domains like fashion. But it also makes things more complicated, introducing new sources of statistical bias and challenges for traditional approaches to training and evaluating algorithms. Humans and machines can also disagree, further complicating the design of production systems. In this talk I'll share lessons from combining algorithms and human judgement for personal styling recommendations at Stitch Fix, an online personal styling service that commits committed to its recommendations through the physical delivery of merchandise to clients.

CCS CONCEPTS

• Ranking • Recommender systems

KEYWORDS

Humans-in-the-Loop

ACM Reference format:

Brad Klingenberg. 2019. **Humans, Machines and Disagreement:** Lessons from Production. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3317086

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317086

What we talk about when we talk about crowdsourcing

A Historical Journey

Maria Stone†

Apple Inc., Cupertino, CA USA maria_stone@apple.com

ABSTRACT

To understand why and how subjectivity and disagreement in label collection matters or doesn't matter I examine the history of systems evaluations and measurement performed by humans and trace the roots of human computation/crowdsourcing and the context in which it arose. Before we can begin fruitful discussions about subjectivity and disagreement we need to ask ourselves what/who it is that human raters are supposed to represent. I offer multiple different perspectives and scenarios that showcase just how varied and ill-defined the role of a human rater can be. I will conclude with some practical recommendations with respect to the questions researchers and practitioners ought to ask themselves before employing human raters, and some challenges with both methodology of such data collection and subsequent analysis of such data.

Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions

Lora Aroyo
Google
New York, NY
l.m.aroyo@gmail.com

Lucas Dixon
Jigsaw
Paris, France
ldixon@google.com

Olivia Redfield
Jigsaw
New York, NY
orhinehart@google.com

Rachel Rosen
Jigsaw
New York, NY
rachelrosen@google.com

Nithum Thain
Jigsaw
Montreal, Canada
nthain@google.com

ABSTRACT

Discussing things you care about can be difficult, especially via online platforms, where sharing your opinion leaves you open to the real and immediate threats of abuse and harassment. Due to these threats, people stop expressing themselves and give up on seeking different opinions. Recent research efforts focus on examining the strengths and weaknesses (e.g. potential unintended biases) of using machine learning as a support tool to facilitate safe space for online discussions; for example, through detecting various types of negative online behaviors such as hate speech, online harassment, or cyberbullying. Typically, these efforts build upon sentiment analysis or spam detection in text. However, the toxicity of the language could be a strong indicator for the intensity of the negative behavior. In this paper, we study the topic of toxicity in online conversations by addressing the problems of subjectivity, bias, and ambiguity inherent in this task. We start with an analysis of the characteristics of subjective assessment tasks (e.g. relevance judgment, toxicity judgment, sentiment assessment, etc). Whether we perceive something as relevant or as toxic can be influenced by almost infinite amounts of prior or current context, e.g. culture, background, experiences, education, etc. We survey recent work that tries to understand this phenomenon, and we outline a number of open questions and challenges which shape the research perspectives in this multi-disciplinary field.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

ACM proceedings, crowdsourcing, subjectivity, toxicity

ACM Reference Format:
Lora Aroyo, Lucas Dixon, Olivia Redfield, Rachel Rosen, and Nithum Thain. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317083

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317083

1 TOXICITY DETECTION

Online discussion forums currently provide a platform for many hazards, such as spreading fake news or expressing personal attacks in the form of online harassment or cyberbullying, typically identified by the use of toxic language in online discussions. Toxicity is a common term used to define negative online interactions in online spaces like Twitter[1], Wikipedia[2], and Google[3]. While the definition is often implied based on context, recent work has formalized one definition of toxicity as "comments that are rude, disrespectful or otherwise likely to make someone leave a discussion" (see, e.g. [21]).

A 2014 Pew Report[4] highlights that 73 percent of adult Internet users have seen someone being harassed online, and 40 percent have personally experienced it. The report outlines two categories of online harassment: (1) name-calling and embarrassment; this is typically considered less severe, however still annoying and often easier to ignore; (2) physical threats and harassment over a sustained period of time, such as stalking and sexual harassment; this category of harassment typically targets a smaller segment of the online population and involves more severe experiences. At the same time, the Wikimedia foundation found that 54 percent of those who had experienced online harassment expressed decreased online participation [47].

The prevalence and impact of this kind of harassment has triggered both the industrial and research communities in the recent years to focus on efficient computational models for online toxicity detection, specifically targeting detection of toxicity in online comments. However, steps in this direction are still in their infancy. Advances in machine learning triggered the development of a number of deep learning approaches displaying very promising performance so far. For example, [16] aims to discover toxic

[1]https://www.wired.com/story/twitter-let-users-hide-replies-fight-toxic-comments/
[2]https://arstechnica.com/information-technology/2017/02/one-third-of-personal-attacks-on-wikipedia-come-from-active-editors/
[3]https://www.perspectiveapi.com/
[4]http://www.pewinternet.org/2014/10/22/online-harassment/

comments in a large pool of documents provided by a current Kaggle competition regarding Wikipedia talk page edits. The authors compare CNNs against the traditional bag-of-words approach for text analysis combined with algorithms proven to be very effective in text classification. The results provide evidence that CNNs enhance toxic comment classification, reinforcing research interest in this direction. Another research initiative, founded by Jigsaw and Google, is Conversation AI[5]. The team is working on tools to help improve online conversation by researching how machine learning methods might help online conversations, what aspects of a conversation machine learning can understand, and what the risks and challenges of using machine learning to assist online conversations are. One specific area of focus in this context is the study of negative online behaviors, like toxic comments. The team has built a range of publicly available models served through the Perspective API[6], including a model that identifies whether a comment could be perceived as "toxic" according to the definition introduced above.

In this paper, we share our initial observations from surveying the current state of the art on detecting negative online behavior and we outline a number of open challenges related specifically to detecting toxicity:

- How does detecting toxicity relate to current approaches for detecting other forms of negative online behavior?
- How do bias and subjectivity impact toxicity assessment?
- How can we account for bias and subjectivity when designing crowdsourcing tasks for toxicity assessment?
- How can we ensure reproducible training and evaluation data for detecting and measuring toxicity in text?

The main contribution of this paper is providing input for a discussion on how we can move beyond the mere detection of sentiment or hate speech to a reliable assessment and ranking of toxicity by better understanding the influence of different personal, societal, and cultural perspectives. We focus the survey on related work on bias and subjectivity in human annotation and how these affect annotation of toxicity in text (section 1.1). We also dive into the current state of the art in detecting negative online behavior and identify the position of toxicity detection there (section 1.2). We position these in the larger context of crowdsourcing subjective assessments (section 1.3). Finally, we outline a number of challenges, open questions, and hypotheses related to measuring rater quality in such subjective tasks (section 2.1), evaluating the performance and reliability of different rating methods for collecting subjective assessments, especially the comparison of absolute rating vs. relative ranking (section 2.2), ensuring the reproducibility of results for toxicity annotation (section 2.3), learning how to rank from sample data (section 2.4), and identifying optimal task designs for subjective assessment tasks (section 2.5).

1.1 Bias and Subjectivity in Human Assessment

Psychologists Tversky and Kahnemann [42] first introduce the idea that judgments, especially subjective ones, can be prone to systematic human biases. Since then, this idea has been applied in various subjective assessment tasks. For example, [1] studies *preference bias* in the context of media recommender systems, showing the effect of anchoring on preference construction. [51] shows that listening tests designed to measure audio quality are vulnerable to systematic errors due to biases with respect to personal preferences, expectations, or mood. [12] exemplifies the "significant detrimental effects" of *cognitive biases* on the annotation quality in relevance judgment tasks, when such biases are not taken into consideration in the design of the task.

To ensure the quality of crowdsourced annotations, there has been a focus on reducing spam produced by bad actors, often by analyzing disagreement. However, spam is typically not the only source of noise in data– subjectivity and unintentional bias can cause disagreement among annotators (which ultimately may serve as a signal). For the purpose of this work, we consider two main categories of subjectivity:

(1) subjectivity inherent to the topic (ie. topics requiring judgments based on personal preference or experience, where two individuals may simply have differing opinions)
(2) subjectivity conditioned by ambiguity in the task (eg. ambiguity of guidelines, input items)

Both types of subjectivity can surface in a single annotation task. For example, in toxicity annotation, two annotators could have different levels of sensitivity to profanity, depending on their background. One person might consider the phrase "fucking awesome" to be toxic, while another would not; this type of subjectivity falls in category (1). Within the same task, they may encounter input items that are ambiguous or not covered by the guidelines that they have to make judgments on, which falls in category (2). It may be possible to mitigate (2) in task design, while (1) may surface regardless.

While the mainstream crowdsourcing approaches often treat disagreement as "noise" and aim their design to decrease and eliminate disagreement, there is an emerging community of researchers that indicate the importance of studying the disagreement and harnessing it to improve the quality of the crowdsourced data. [4] shows how the established practices in human annotation treat disagreement as "noise" and further [3] introduces the notion of "disagreement-based quality metrics", illustrating how the disagreement in crowdsourcing can be harnessed to achieve more representative and reliable annotation data. [24] proposes another approach for harnessing the disagreement through the *"crowd parting"* method to identify divergent, but valid, worker interpretations in crowdsourced tasks. [46] deals with a different approach to attend to the diversity of human annotators through a multi-dimensional measure of the annotators' ability. [38] demonstrates in the domain of semantic relatedness how the notion of a "universal gold standard" is highly problematic. The authors point out the importance of understanding these gold standard datasets from a human-centered point-of-view. Concretely, they empirically demonstrate that "people who belong to different cultural communities would provide different answers to a variety of knowledge-oriented tasks."

With respect to disagreement-aware task design, the CrowdVerge approach predicts which visual questions will have more disagreement and adds annotators to those questions in order to better capture the diversity of possible answers [19]. [8] explores limiting guidelines in favor of post-hoc label decisions, which reduces training time and surfaces potentially unanticipated ambiguity in

[5]https://conversationai.github.io/
[6]https://perspectiveapi.com

the task through disagreement. We explore open questions around task design further in section 2.5.

1.2 Detection of Negative Online Behaviours

As social media has become an unprecedentedly large and open space for information publishing and discussion, it is only natural that it would also attract forces that aim to exploit and misuse it to spread content that can be degrading, abusive, or otherwise harmful to people. To address this issue, research into automatic detection of different types of negative online behavior has emerged. While toxicity is a popular direction of study, it stands alongside other dimensions of negative online interaction. For example, [47] investigates personal attacks, [25] aggression, [7, 13, 45] hate speech, [29, 49] harassment, and [11] cyberbullying. [50] provides an in-depth overview of research in this area. There is also growing interest in understanding how these attributes relate to one another; for example, [20] investigates the connections between insulting, obscene, threatening, and hateful language and self-reported data on user behaviour.

As [47] points out, much of this work builds on existing machine learning approaches in fields like sentiment analysis [5, 35, 50, 52] and spam detection [37, 39]. However, one of the biggest challenges for detecting any of the forms of negative online behavior is the fact that all of them are influenced and defined by "the prevailing social norms, context, and individual and collective interpretation" [13]. Similarly, as seen in section 1.1, individual annotator differences make it difficult to identify such behavior consistently, and thus annotation tasks tend to result in high disagreement between annotators [27].

In addition to those challenges, it should be noted that even the definition of toxicity we discussed in Section 1 is highly context and community-dependent. What may be considered toxic on one platform or forum may be acceptable on another, due to the diverse nature of discourse in communities online.

1.3 Crowdsourcing Subjective Assessment

From the overview of research on bias and subjectivity (1.1) and how these influence and complicate the detection of negative online behaviours (1.2), we conclude that it is not trivial to ensure the reliability of human-annotated training data. It is critical to understand how each of these tasks inherently carry human bias and subjectivity, and to translate this to an adequate annotation task design.

One component that plays a central role in task design is the method for eliciting ratings. Absolute rating and comparative rating (or some combination of the two) are commonly leveraged. **Absolute rating** (often referred to as *judgment on a scale*) commonly asks annotators to rate the absolute value of some trait in the provided datum. There is extensive research ([30], [28], [9], [15], [18]) on different types of scales (e.g. Gunman scale, Likert scale at 7-point, 5-point, 3-point, or binary assessment, etc.) used to collect and measure user input, e.g. *nominal variables*: label a series of values; *ordinal scales*: order of values; *interval scales*: order of values and the ability to quantify the difference between each one; *ratio scales*: order, interval values and the ability to calculate ratios. **Comparative rating**, often referred to as *relative judgments*, frames the rating task as *a comparison*: asking raters to compare two items on a single characteristic; or in a *multiple comparison* setting, placing a number of items in order according to a single characteristic [23], [10]; [32], [14], [41], [48]. The research results of [17] support the hypothesis that *"humans may often be able to make more accurate ratings using comparative measures"* [26], [40], [34], [33].

We propose **hypothesis 1** that, especially in the case of toxicity assessments, agreement will differ between ratings produced by relative rating as opposed to an absolute scale. That is to say, we imagine it will be easier for raters to agree on a set of paired comparisons than it would be for them to agree on whether individual comments should be considered toxic.

Following the challenge of choosing the most appropriate assessment scale is that of generating a plausible rank from the collected judgments. In order to allow transferability of the results and their reuse within different ML systems, it is important to be able to convert results from one scale to another. In this process, however, it is critical to understand the loss of accuracy or meaning across scales. It could also be interesting to examine how raters' quality varies in these transformations, and also how we evaluate the accuracy of the transformation. When transforming either the absolute or relative ratings to linear rank, typically the distances between different points on the scale may not map uniformly.

2 CHALLENGES & OPEN QUESTIONS

Current models for capturing and understanding negative online behaviors, just like all machine learning models, still make errors. However, the challenges go beyond just improving the accuracy in detection. It is critical to provide more interactivity, transparency, and granularity for end users to be able to understand the range and diversity of the toxicity types (e.g. being able to select which types of toxicity they are interested in finding) and how they relate to specific context and cultures.

Thus, one of the grand challenges for such machine learning models is to understand the variety of human subjectivity and implicit (unintentional) bias. This is especially important when gathering training data for machine learning. One step towards solving this problem is the awareness that some data collection tasks carry intrinsic human subjectivity, and this needs to be understood in order to provide reliable training data. Research indicates that subjective assessment tasks are highly prone to disagreement between judges [43, 44]. This is typically caused by the ambiguity of the text and the inherent subjectivity of human raters. For example, Alonso and Mizzaro [2] found that the crowd relevance labels contradict the labels given by experts. However, individual worker judgments are still combined, without accounting for disagreement and ambiguity, by using majority vote or expectation maximization algorithms [31].

Below, we list a number of questions that deal with the challenges related to subjective task assessments.

2.1 Measuring rater quality in subjective tasks

As one of the goals of subjective task assessments is to gather the full range of human opinions, perspectives, and interpretations, inter-rater agreement is not an adequate measure for quality [4,

24]. We believe that metrics harnessing the diversity each rater contributes are needed in this context. However, the challenge is to distinguish between 'good' disagreement and 'bad' disagreement, which could be the result of spamming behavior [39]. We propose the following **hypothesis 2**: the distribution of raters disagreement will be indicative of raters perspectives on judging the toxicity of the text. In other words, the number of groups of ratings with similar counts can indicate the number of cultural or social perspectives with respect to the toxicity in the text. An additional **hypothesis 3** can be made - high disagreement in rater votes can indicate high ambiguity in the language used.

2.2 Evaluating different rating methods

To compare variants of the Revolt annotation approach with traditional crowdsourcing methods, [8] ran a series of annotation experiments, running each condition on eight different tasks with different datasets, ensuring no annotator saw more than one condition for a single dataset, and measuring the accuracy of the resulting annotations. Measuring effectiveness in subjective tasks can prove very challenging, as typically these tasks have no clear ground truth, or even if there is some notion of it, it can fluctuate over time. Thus, new metrics for measuring success and quality of the collected data need to be applied to capture or discriminate between different influences, e.g. time, culture, etc. A significant body of research focuses on how people rate differently on these scales, outlining the pros and cons of different scales (e.g. relative ratings vs. linear scores; Likert scales on different points compare; symmetric vs. asymmetric scales, etc) and how this impacts the reliability of the results.

Below we outline a number of observations and hypotheses with respect to the comparison of **relative rating vs linear scores**:

Severity of Toxicity: When rating a single comment, raters lack context about the full spectrum of toxicity among comments. As a result, a rater might rate a petty insult with the highest toxicity point on the scale without knowing that another comment in the pool contains more severe language, such as obscenities or threats. For example, when displayed independently, the comments "you are stupid" and "you should die" are both unambiguously toxic; consequently, they would likely be ranked the same in a single comment rating scheme, and models trained on this data would not contain any notion of severity. However, when displayed side by side, the difference between these comments becomes clearer, and the ratings are more likely to reflect that. Having raters understand this distinction is important because during an in-person conversation, the consequences of saying the threatening statement would be much harsher than those of saying the insulting one, and users often expect that models of online conversations will reflect these dynamics. We propose the following **hypothesis 4**: If a rating task requires rating a pair of comments relative to one another on a scale, then the overall comment rankings will more accurately reflect the severity of toxicity than the corresponding single comment task.

Keyword spotting: Certain words have strong connotations associated with them, and the perceived sentiment of the presence of certain language can have a strong effect on the perception of toxicity. However, the flexibility of language allows subtle changes in sentence structure to completely change the meaning of a phrase containing a specific word; examples include negation, quotation, prepositional phrases, and slang, among others. Consider the difference between the phrases "Go to hell" and "Hell yeah!", or between "Please don't hurt yourself" and "Please hurt yourself"; both sets of phrases contain similar words, but the toxicity of the comments differs completely. If seen in isolation, raters might only see the negative words in these phrases, but when asked to compare a pair, they may examine them more closely and find the distinction. We present the following **hypothesis 5**: If a relative rating task requires rating a pair of comments relative to one another on a scale, then there will be fewer errors due to "keyword spotting", which we define as a rater looking at words in isolation to make a judgment of toxicity, rather than the overall sentence.

Rating Variance: When presented with a comment with mild toxicity (e.g. a petty insult) in a pair of comments to compare, a rater might feel that the comment in question is considerably more toxic than the other comment (e.g. a friendly greeting), but given a separate comment to compare it to (e.g. a threat), the rater might feel that the comment in question is considerably less toxic than the other. This situation would yield two vastly different ratings on the Likert scale for two pairs containing the same comment, which would result in higher variance of ratings for that comment. Alternatively, when presented with a comment with extreme toxicity in a pair, the rater is likely to always say that the extremely toxic comment is much worse than any other comment it is compared to; this would yield lower variance among the ratings. With a single comment scale, all ratings for both the mildly toxic comment and the more extremely toxic comment are both likely to be on the higher end of the scale, since there is no comparison to be made; variance would be low in both cases. We present the following **hypothesis 6**: Relative rating variances of a comment considered across all pairs it could appear in should be higher than the absolute rating for the same comment, due to the range of contexts of the former (i.e. most comments appear alongside better and worse comments). The only comments with low relative rating variance should be those that are very obviously toxic (maxima in our ranking) or obviously not toxic, though the latter usually tends to be more ambiguous.

2.3 Ensuring reproducibility of results

It is important that human computation tasks be reproducible [36], but there are a number of bottlenecks related to reproducibility of results. One of the main challenges is the fact that individual annotation efforts are performed at a specific moment of time, while language, cultural context, and information evolve in the meantime. Each of these factors introduces a new perspective on the way text can be interpreted in a different period in time. Evolution of language online can be observed by studying social media data [22].

For our purposes, semantic change can be classified in two ways: **pejoration**, which is when a "word is used to express negatively loaded values not inherent in its historically original (or historically prior) meaning" and its opposite, **amelioration** [6]. Pejoration occurs more frequently and is therefore more relevant to the

classification of toxicity [6]; instances of pejoration of language introduce two potential causes of inconsistency into annotations of comments containing these newly toxic words:

(1) A rater more familiar with a newer, toxic context of a word could mistakenly rate a comment from years prior containing the word in a non-toxic context as toxic.
(2) A rater not familiar with the newer, toxic context of a phrase could mistakenly rate a recent toxic comment as non-toxic.

Corollaries to these cases would involve the same scenarios with amelioration of language.

Case (1) will inhibit reproducibility because the same experiment performed years later could yield false positives, contradicting the original negative annotations. Case (2) has a similar issue; if the experiment is repeated after the toxic phrase has become more well known, then the newer annotation will contradict the earlier, falsely negative annotation.

While it may inhibit reproducibility initially, case (2) will produce more accurate experimental results if the experiment continues to be repeated. Conversely, repetitions of case (1) will cause further divergence from the original results over time.

hypothesis 7: we hypothesize here that annotation efforts need to be performed continuously, so that annotated datasets can evolve with the latest linguistic, cultural and social contexts. Additionally, we hypothesize that capturing temporal snapshots of annotation efforts allows for a temporal analysis of the toxicity evolution with respect to severity and variance (as noted in section 2.2).

For instance, if we were to reproduce the crowdsourcing results from [47], using the same set of comments, same instructions, and same annotation platform, it would be important in our analysis to consider the temporal factor, where semantic change and cultural context may alter the results, as well as differences in annotator backgrounds.

2.4 Learning a rank from sample data

While collecting toxicity assessments at various scales, ultimately we aim at producing a ranked list of comments. The process of transforming the ratings at scale to a representative rank is not trivial for a number of reasons. We outline a few **open questions** that are important to consider in this context, e.g. what is an optimal amount of data to learn a rank; how well do probabilities correspond to severity; how well distributed are the values (linearization of the relative scores) vs the pre-selection methods based on scores from probabilities; how do sampling methods influence the optimal learning, and how to incorporate active learning in the process of collecting rating data?

2.5 Task Design for Subjective Tasks

In previous sections we hypothesized that disagreement is an indication of the differing annotator interpretations/perspectives and potential ambiguity in the text. In order to allow for this disagreement to be expressed and to enable its harnessing, the crowdsourcing task design should incorporate a number of elements. Thus, our **hypothesis 8** here is that in order to achieve an optimal design setting, a number of pilot experiments are needed with the following design considerations:

- what is an adequate *assessment scale* to deal with the variation and inconsistency of human judgments, both within and between subjects (see 2.2)
- what is an adequate *task template* that enables capturing the full spectrum of opinions, i.e. aiming at diversity in answers vs. optimizing for consensus (as indicated in sections 2.1 and 2.3)
- what is the optimal *number of raters* that will allow us to gather reliable and reproducible data; we hypothesize here that a higher number of raters will help to capture the full spectrum of diversity in opinions for tasks that are highly subjective or with highly ambiguous text.

ACKNOWLEDGMENTS

We acknowledge the contributions of our collaborators Chris Welty and Marie Pellat. We thank Lucy Vasserman and the workshop reviewers for their feedback.

REFERENCES

[1] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* 24, 4 (2013), 956–975. https://doi.org/10.1287/isre.2013.0497 arXiv:https://doi.org/10.1287/isre.2013.0497
[2] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *IPM* 48, 6 (2012), 1053–1066.
[3] Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1. https://doi.org/10.15346/hc.v1i1.3
[4] Lora Aroyo and Chris Welty. 2015. Truth is a Lie: 7 Myths about Human Annotation. *AI Magazine* 36, 1 (2015).
[5] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. *CoRR* abs/1709.04219 (2017). arXiv:1709.04219 http://arxiv.org/abs/1709.04219
[6] Paulina Borkowska and G. Kleparski. 2007. It befalls words to fall down: pejoration as a type of semantic change. *Studia Anglica Resoviensia* 47, 4 (2007), 33–50.
[7] Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making.. In *Proceedings of Internet, Policy and Politics Conference*. Oxford, United Kingdom.
[8] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI âĂŹ17*. ACM Press, 2334âĂŞ2346. https://doi.org/10.1145/3025453.3026044
[9] Dennis L Clason and Thomas J Dormody. 1994. Analyzing data measured by individual Likert-type items. *Journal of agricultural education* 35 (1994), 4.
[10] Herbet Aron David. 1963. The method of paired comparisons. (1963).
[11] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 18 (Sept. 2012), 30 pages. https://doi.org/10.1145/2362394.2362400
[12] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654
[13] Anirudh Srinivasan Adam Glynn Jacob Eisenstein Eshwar Chandrasekharan, Umashanthi Pavalanathan and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 1. ACM, New York, NY, USA, 22. https://doi.org/10.1145/3134666
[14] V.V. Fedorov. 1972. Theory of optimal experiments. *Elsevier* (1972).
[15] Gerhard H Fischer and Ivo W Molenaar. 2012. *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.
[16] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN '18)*. ACM, New York, NY, USA, Article 35, 6 pages. https://doi.org/10.1145/3200947.3208069
[17] Richard D. Goffin and James M. Olson. 2011. Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others. *Perspectives on Psychological Science* 6, 1 (2011), 48–60. http://www.jstor.org/stable/41613423

[18] Joy Paul Guilford. 1954. Psychometric methods. (1954).

[19] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI âĂŹ17*. ACM Press, 3511âĂŞ3522. https://doi.org/10.1145/3025453.302578

[20] Toby Hopp, Chris J Vargo, Lucas Dixon, and Nithum Thain. 2018. Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept. *Social Science Computer Review* (2018), 0894439318814241.

[21] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving GoogleâĂŹs Perspective API Built for Detecting Toxic Comments. *arXiv preprint arXiv:1702.08138* (2017).

[22] Noah A. Smith Jacob Eisenstein, Brendan O'Connor and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9, 11 (2014).

[23] Shaili Jain Anil K. Jain Jinfeng Yi, Rong Jin. 2013. Inferring UsersâĂŹ Preferences from Crowdsourced Pairwise Comparisons: A Matrix Completion Approach. *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* (01 2013), 207–215.

[24] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. 1637âĂŞ1648.

[25] Joseph M Kayany. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on Usenet. *Journal of the American Society for Information Science* 49, 12 (1998), 1135–1141.

[26] Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and BestâĂŞWorst Scaling. (2016).

[27] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI'13)*. AAAI Press, 1621–1622. http://dl.acm.org/citation.cfm?id=2891460.2891697

[28] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[29] Melanie J Martin. 2002. Annotating flames in Usenet newsgroups: a corpus study. *For NSF Minority Institution Infrastructure Grant Site Visit to NMSU CS department* (2002).

[30] Michael S Matell and Jacob Jacoby. 1971. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and psychological measurement* 31, 3 (1971), 657–674.

[31] Tyler McDonnell, Matthew Lease, Tamer Elsayad, and Mucahid Kutlu. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *HCOMP*. 10.

[32] Sir Moser, Claus and Graham Kalton. 1972. *Survey methods in social investigation* (2nd american ed ed.). New York : Basic Books. Previous ed.: 1958.

[33] Jack O'Neill, Sarah Delany, and Brian Mac Namee. 2018. From Rankings to Ratings: Rank Scoring via Active Learning. (10 2018).

[34] Jack O'Neill, Sarah Jane Delany, and Brian Mac Namee. 2017. Rating by Ranking: An Improved Scale for Judgement-Based Labels. In *IntRS@RecSys*.

[35] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. https://doi.org/10.1561/1500000011

[36] Praveen Paritosh. 2012. Human Computation Must Be Reproducible. In *CrowdSearch: WWW Workshop on Crowdsourcing Web Search*. 20–25.

[37] Vikas C. Raykar and Shipeng Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13 (Feb. 2012), 491–518. http://dl.acm.org/citation.cfm?id=2188385.2188401

[38] Rebecca Gold Benjamin Hillmann Matt Lesicko Samuel Naden Jesse Russell Zixiao (Ken) Wang Shilad Sen, Margaret E. Giesel and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW âĂŹ15)*. ACM, New York, NY, USA, 826âĂŞ838. http://dx.doi.org/10.1145/2675133.2675285

[39] Nikita Spirin and Jiawei Han. 2012. Survey on Web Spam Detection: Principles and Algorithms. *SIGKDD Explor. Newsl.* 13, 2 (May 2012), 50–64. https://doi.org/10.1145/2207243.2207252

[40] Saif M. Mohammad Svetlana Kiritchenko. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. (2017).

[41] L. L. Thurstone. 1927. A Law of Comparative Judgment. (1927).

[42] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:http://science.sciencemag.org/content/185/4157/1124.full.pdf

[43] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IPM* 36, 5 (2000), 697–716.

[44] Ellen M Voorhees. 2001. The philosophy of information retrieval evaluation. In *CLEF*, Vol. 1. Springer, 355–370.

[45] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.

[46] P. Welinder and P. Perona. 2010. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference*. 25âĂŞ32. http://dx.doi.org/10.1109/CVPRW.2010.5543189

[47] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on the World Wide Web*. Perth, 1391–1399.

[48] Bernardo A. Huberman Yarun Luon, Christina Aperjis. 2012. Rankr: A Mobile System for Crowdsourcing Opinions. (2012).

[49] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[50] Justine Zhang, Jonathan P Chang, Lucas Dixon, Yiqing Hua, Nithum Tahin, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1.

[51] Slawomir Zielinski, Francis Rumsey, and SÃ¿ren Bech. 2008. On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review. *J. Audio Eng. Soc* 56, 6 (2008), 427–451. http://www.aes.org/e-lib/browse.cfm?elib=14393

[52] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Manage. Inf. Syst.* 9, 2, Article 5 (Aug. 2018), 29 pages. https://doi.org/10.1145/3185045

Discovering User Bias in Ordinal Voting Systems

Alyssa Lees
Google Research
alyssalees@google.com

Chris Welty
Google Research
cawelty@gmail.com

ABSTRACT

Crowdsourcing systems increasingly rely on users to provide more subjective ground truth for intelligent systems - e.g. ratings, aspect of quality and perspectives on how expensive or lively a place feels, etc. We focus on the ubiquitous implementation of online user ordinal voting (e.g 1-5, 1 star-4 stars) on some aspect of an entity, to extract a *relative* truth, measured by a selected metric such as vote plurality or mean. We argue that this methodology can aggregate results that yield little information to the end user. In particular, ordinal user rankings often converge to a indistinguishable rating. This is demonstrated by the trend in certain cities for the majority of restaurants to all have a 4 star rating. Similarly, the rating of an establishment can be significantly affected by a few users [10]. User bias in voting is not spam, but rather a preference that can be harnessed to provide more information to users. We explore notions of both global skew and user bias. Leveraging these bias and preference concepts, the paper suggests explicit models for better personalization and more informative ratings.

ACM Reference Format:
Alyssa Lees and Chris Welty. 2019. Discovering User Bias in Ordinal Voting Systems. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3317080

1 INTRODUCTION

Crowdsourcing systems increasingly rely on users to provide more subjective ground truth for intelligent systems - e.g. ratings, perspectives on how expensive or lively a place feels, etc. Not surprisingly, there are no questions for which a pure consensus can be reached through on-line voting. Disagreement is the normal case, and the important issue is how to most productively process that disagreement into a usable signal. In this paper we examine the problem of how to process *ordinal votes*, i.e. votes on a fixed scale (e.g. 1-4 or 1-5), with Likert properties[8], from multiple users on different subjective aspects of a rated entity. Likert[1] referred to the human responses to subjective questions as *attitudes*.

We begin with a few observations of ordinal systems in practice. There are many publicly available systems that provide results of ordinal voting, such as star ratings, of movies, restaurants, music, etc:

- Mean ratings tend to converge: many rated things (movies, restaurants, songs) have 4 out of 5 stars.
- In use-cases where fractional ratings are displayed based on ordinal votes, discrepancies between ratings in the same range yield little information in terms of differentiation of quality.
- There is user bias: some users consistently vote higher or lower than others and these votes can be affected by underlying preference
- A single entity ranking can be significantly affected by a few users [10].
- Votes are *subjective*, they reflect not only attitudes, but contextual information that may or may not be accessible when processing the votes
- *There is no ground truth*

Our research question is derived from these observations:

> **RQ1**: Can we identify *consistent* user voting bias and productively harness its signal for more reliable and/or informative ratings?

The standard way to process ordinal votes is to take the majority, or plurality (the choice that receives the most votes), or some other aggregate, most commonly the mean. A naive approach standard way to deal with user bias is to throw out votes that disagree with the majority, or to down-weight them as a function of their overall agreement. Any methodology expanding on this approach is performing a form of outlier detection and will devolve in to the all 4-star restaurants problem addressed above. It is worth noting, that many existing systems have expanded on these notions to learn more elaborate user *trust* scores [2]. User trust is an integral element of evaluating any online voting system, but only represents one dimension.

In this position paper we outline an approach for identifying different kinds of bias, from users as well as the rated entities themselves, and how it can be productively harnessed. If user voting patterns are consistent, we should be able to map them into a usable signal – if someone consistently votes lower than everyone else, in theory it should be possible to re-normalize their voting behavior and turn disagreement into agreement.

Finally, one of the largest sources of voter disagreement may be attributed not to explicit behavioral bias, trust or expertise, but rather user preference. Distinguishing preference, i.e. preferring one type of movie over another, is integral to personalization of systems as well as finding a system where all global ratings don't devolve to a standard value.

Aligning biased user votes can be important in a number of ways, not least of which is increasing the number of usable votes that are available at places for which there are few votes already. For example, users probably don't need to know what other users think of McDonald's or Star Wars, there is not much we could

[1] pronounced lick-ert [7]

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317080

learn, but some small out-of-the-way shop or Indie movie that has three recorded votes may turn out to be a gem, if only we could understand that seemingly divergent votes actually agree.

2 BACKGROUND

Identifying consistent voting behaviors can be considered a generalization of the Bayesian Truth Serum [12], which cleverly identifies a particular voting pattern among experts, in cases where expertise applies, such as the famous wine tasting game; experts tend to disagree with the crowd, but are also capable of predicting that disagreement.

In the case of star-ratings, one could argue that some amount of expertise and background knowledge applies, but for more quantitative tasks like rating the price level of a restaurant, expertise is less obvious. We could consider the range of places a person has visited as expertise. In general, we simply search for consistent, patterns of voting behavior by comparing users to some relative truth, such as the plurality vote, the mean or as suggested in this paper, an informed prior.

Much recent work on bias in machine learning looks for different kinds of bias, bias that favors one population of users in ways we may consider unfair, such as race, gender, etc [4]. However, in this work we focus on explicit differences in user voting behavior only. These differences can be roughly divided into subcategories of voting behavioral distribution trends, preference and reliability/trust.

Existing work on user reliability in the context of online social media usage and online question answer systems [2, 3], and user voting behavior bias [18] suggest a re-ranking methodology to remove the existence of these behavioral data biases. Explicit user behavioral preference in online queries can be exploited for better click-through performance [19].

There is a wealth of work in personalization, e.g. via Bayesian probabilistic models extracted from learned explicit user preferences for recommendation systems [15]. Online aspect ratings can be predicted from user reviews [16] as well as user data in ordinal voting systems [11]. The focus on *aspects* of the places being voted on, such as ambience, food quality, etc. in [11] explicitly analyzed the bias of user populations via an Ordinal Aspect Bias Model using Bayesian inference.

While this paper will focus on a high level model of different user biases in ordinal voting systems, we are inspired by the rich background of collaborative filtering techniques that assist in clustering and identifying behavioral patterns and rating prediction [5, 9]. It is important to note that despite the nature of the ordinal votes used in these techniques, there has been an underlying assumption that the rankings follow a continuous valued distribution. We will continue that assumption in our bias discussion.

Many statistical techniques utilizing Bayesian inference require cutoff points in a Gaussian distribution, other work uses a logitic regression model for category data [6]. Another alternative is the logistic stick-breaking process for separating clusters and mapping the categorical likelihood [14].

Finally, there has been working expanding existing Bayesian co-clustering techniques [17] for identifying user behavior preferences. This work has exploited Gibbs sampling [1] to generate examples from the posterior for improved parameter estimation.

3 VOTING BIAS IN ORDINAL SYSTEMS

In order to process votes to identify user bias, we make a few assumptions. Given a set of rated entities, $e \in \mathcal{E}$, users $u \in \mathcal{U}$, and ordinal user ratings $r_{u,e} \in \mathbb{N}^{[0,M]} = \{x \in \mathbb{N} | 0 \leq x \leq M\}$, where M denotes the maximum ordinal voting value. We denote the *true rating*, independent of bias, of a rated entity in a continuous domain as $v_e \in \mathbb{R}$. Although we have no expectation to ever know any particular v_e, we do make assumptions on the prior distribution from which any particular v is drawn. For simplicity, we model the prior with a Gaussian normal distribution, $P(\mathcal{V} = v) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ that is truncated at the domain value cutoffs.

Next we allow for an aggregation function on the set of votes for an entity, to produce a set of raw ratings $r_e \in \mathbb{R}$ from the set of user ratings, $r_e = \mathfrak{F}_e([r_{u,e}])$. Typical choices for the function $\mathfrak{F}$ are the mean or plurality (most popular ordinal vote). Since individual votes are subject to bias, the distribution of raw ratings will be skewed by this bias, and will have a different, possibly non-Gaussian, distribution.

In the simplest case, and the standard approach to disagreement, users vary consistently in their reliability, so that there is a user trust vector $\bar{t}$, such that $v_e = \mathfrak{F}'_e([r_{u,e}], [t_u])$. Typical choices for $\mathfrak{F}'$ are weighted mean, or weighted plurality.

Mitigating the ineffectual information relayed by such votes can begin simply by filtering outliers (both in voting and users), changing how distributions of votes are displayed, re-weighting and catering to specific user preferences.

We explore several avenues towards imparting more informative user votes:

- Explore explicit aspect voting bias to identify common voting behaviors and preferences via clustering.
- Identify user zeitgeist behavior (bias) to inform global distributions of votes
- Utilize location bias (say for restaurants and hotels) to change conditional distributions of votes.
- Identify individual user preference (bias) and leverage for personalization.
- Identify individual voting behaviors and normalize bias-independent display
- Leverage user trust scores for increased accuracy in ranking.

4 BIAS/PREFERENCE MODEL

We propose modelling user ordinal voting systems with an assumption of a theoretical *true* underlying prior probability distribution function for the ranking of the entities E. In this idealized model, individual unbiased *true* entity rankings are drawn from the PDF. There are two not necessarily intuitive points with this assumption: 1. the underlying entity rankings are drawn from continuous distribution 2. that de-biased entity model may reflect a different distribution than the ordinal voting samples.

For a subjective voting system, such as movie ratings, a suitable prior may be a normal distribution. While unbiased selections should naturally follow a bell-curve, with the majority ranking average and a few achieving excellence, a more objective voting task, such as one requesting users to rate the price level of a restaurant, may be informed by the real-life skew of establishments. In sample data sets across many regions, the authors found that restaurant

data demonstrates a skew towards inexpensive and moderate establishments. In this scenario, a suitable prior could be modelled by a Gamma distribution.

We extend our definition of the *true* rating of an entity to cover this expected distribution:

$$v_e : e \in \mathcal{E}, P(\mathcal{V} = v) = F_V(v)$$

where V is a continuous random variable with a prior modelled by the function $F_V(v)$. Again, this *true* value is not necessarily observed in sampled user votes and likewise is not necessarily representative of what should ideally be displayed in any online ranking/rating system for the entities $\mathcal{E}$. Rather we argue that these continuous values are skewed by several kinds of bias. The first kind of bias, which we will deal with later, is user bias. The second kind of bias, which we discuss in this section, is a natural bias caused by extrinsic and global factors, that we consider properties of the voting question. Examples of such natural bias include culture, geographic location and genre, and there are many more.

Location bias includes the notion of rural versus urban; a rural area may apply different standards of expectations for price and menu items than a more populated area. Similarly, different countries and regions have demonstrated clear differences in dominant preferences and cultural tastes. The concept of location applies to search query rankings, movie rankings, and almost any conceivable online ranked result.

Genre bias includes particular styles for a given entity class. In the space of movies, genre may encompass *action* or *romantic comedy*. In the space of restaurants, genre could be a type of cuisine such as *Italian* as well as experiential features of the establishment such as *trendy, casual,* or *upscale*.

Genre and location are often intertwined. For example the genre of Bollywood movies are more popular in specific regions. Likewise the skew of ranking may be similarly affected.

4.1 Natural Bias

These notions of **Natural Bias** where for particular combinations of e.g. location and genre, the zeitgeist opinion of users or the actual entities themselves demonstrate a modified distribution from the underlying F_V prior. The point here is that natural bias may distort and skew the distribution of votes according to trends that are separable from individual user bias. Action movies tend to rate higher, they have their own distribution, rural areas have different price level standards, and its hard to find a genuinely expensive restaurant.

To account for these natural biases we define sets to partition the entities, and distributions within those sets. For example, we define genre $g_e \in \mathcal{G}$ as the genre of entity $e \in \mathcal{E}$, and $l_e \in \mathcal{L}$ as the location of entity e. Note that in the context of user votes, the location of the user voting l_u would be a separate consideration, but is outside the scope of this paper.

We call the distribution of the rated entities with their natural bias score $Z_{g,l}$ and we define $z_e \in Z$ to be the natural bias adjusted rating of an entity. Again, we don't actually have these values, but we expect the overall distribution to be closer to the distribution of entities by their raw votes, $r_e \in \mathcal{R}$. We define a function $h : \mathbb{R} \mapsto \mathbb{R}, (v, g, l)$ that reflects the mapping between the ideal distribution and the natural bias, such that:

$$P(z_e|V, G, L) \sim h(v_e, g_e, l_e)$$

Similarly, we define the inverse, mapping naturally biased entity ratings back to the ideal distribution V as

$$h^{-1}(Z, G, L) \sim V$$

In other words, if we can characterize the bias according to genre and location as distributions, given an ideal distribution we can collect the bias-adjusted rating of individual entities.

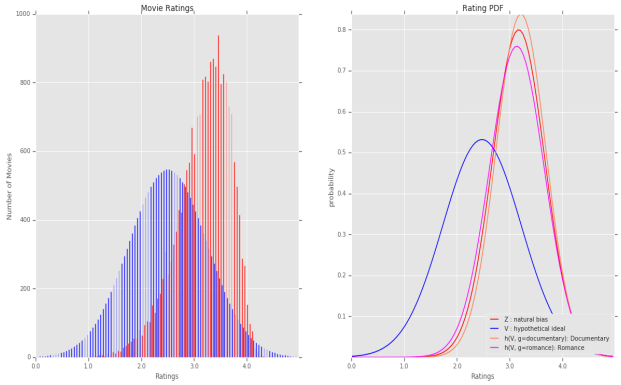


Figure 1: Idealized Movie Rating Distributions with Natural Bias

All we really have, however, is raw user votes on each entity ($R_e = r_{1,e} \ldots r_{n,e}$), where each vote is on the ordinal scale, leading to some raw distribution of entities on a continuous, real-valued scale. Our next problem is converting the scale of those distributions back into the ordinal scale. If, for example, the entities follow a normal distribution centered on 1.4, at what point do we "cut off" and say an entity has a 2, or a 1, rating?

Rating Likelihood: Stick Breaking Dirichlet Process. We represent the probability of any given user-entity rating with a specific ordinal vote value as $P(r_e = m)$ where m is an ordinal value $m \in \{1, .., M\}$. There are many ways to model discrete values to a continuous distribution. Here we will follow established precedent to employ the Dirichlet Stick-Breaking process. In this representation, the probabilities $P(r_{u,e} = m)$ are defined by a procedure of cutting up a unit length stick at cut points $C = \{c_1, ..., c_{M-1}\}$ where $c_1 < c_2 < \ldots < c_{M-1}$. The purpose of this cutting is essentially to find the best real-valued points in $[1, M]$ to use as cut points to divide the distribution into the original ordinal scale.

The stick-breaking Dirichlet process itself contains two variables, locations $\{\theta_m\}$ that are independent and identically distributed over the process and the corresponding probabilities $\{\beta_m\}$. The Dirichlet process probability mass distribution for a discrete set of values has a probability mass function:

$$f(\theta) = \sum_{m=1}^{m=\infty} \beta_m * \delta_{\theta_m}(\theta)$$

and an indicator function $\delta_{\theta_m}(\theta_m) = 1$ and zero for every other ordinal voting value. We will need to relate θ to our cut points in C.

The stick breaking process begins by viewing β_m as a length of a piece of a unit stick broken at m and the remaining stick as $\beta_m^{'}$, such that

$$\beta_m = \beta_m^{'} * \prod_{i=1}^{m-1} (1 - \beta_i^{'})$$

In our scenario, $\zeta_e^m = |c_m - z_e|$ is the distance from the idealized rating to the cut point, so that the actual probabilities align, $P(r_e = m|z_e, C) = p(r_e = m|\zeta_e^m)$. To translate to the stick breaking formula, we use the sigmoid function to ensure that the ζ distance from cutoffs is reduced to a unit length stick. As such,

$$\beta_m = s(\zeta_e^m) = \frac{e^{\zeta_e^m}}{1 + e^{\zeta_e^m}}$$

and

$$P(r_e = m) = \prod_{m^{'} < m} (1 - s(\zeta_e^{m^{'}})) * s(\zeta_e^m)$$

Finally, we approximate z_e as the expected value of the individual user ratings:

$$P(Z = z_e) \sim E(r_e) \sim \frac{1}{|U_e|} \sum_{u \in U_e} r_{u,e} \sim h(v_e, g_e, l_e)$$

4.2 User Reliability

As highlighted in the introduction, real life scenarios are riddled with instances where a few user voters can significantly alter the outcome of online voting. Examples include spam, where a user votes randomly or for one value on numerous entities[13]. Alternatively, there are examples where friends of an establishment repeatedly vote online to mitigate negative online rankings. Notable instances have been recorded where a single disgruntled customer can offer a poor rating that throws the displayed ranking of a product shopping sight off for months. This is especially problematic, when ratings translate effectively become binary - i.e. everything above say a 4 start level is about the same and regarded as decent and anything below is viewed with skepticism.

To address, our model of user bias must incorporate an individual user trust score $t_u \in \mathbb{R}^{[0,1]}$ and T_U is the set of trust scores across users in U.

$$P(Z = z_e) \sim \frac{\sum_{u \in U_e} t_u * r_{u,e}}{\sum_{u \in U_e} t_u}$$

Modelling user reliability/trust can be learned via historical data on individual basis. Alternatively suspected outlier detection can be learned more broadly in the user and entity domain spaces.

4.3 User Voting Behavior and Preference Biases

Additionally we assume that users have *behavioral* and *preference* bias that should be treated separately from *trustworthiness*. These types of bias can be used for personalization, recommendation and individualized tailoring of ranking results.

(1) **Voting Behavioral Bias** : We will define voting behavioral bias as $b(u, e)$. Voting behavioral bias is NOT due to user trust and can take many forms.
- Users may vote different if given a series of ordinal questions online. Studies have demonstrated that more attention may be given to the first question then later items.
- Some users have more optimistic or pessimistic trends. In other words, some users may have a bias towards voting one level higher than the norm while others may vote lower. If identified, these voting distributions can be re-normalized for finer grained results.
- Expert versus Novice voting. A novice may conform to more popular voting conventions while an expert may display a wider range and finer grained discrepancy between votes.

(2) **User Preference** : Individual users display different individualized preferences dictated by genre and location: $a(u, g_e, l_e)$. These individuals preferences are separate from the Global Bias trends due to these factors. For example, a user may prefer action movies or romantic comedies, and hence are prone to giving higher rankings to *Action* movies independent of underlying movie quality. The same is applicable to restaurants, shopping etc. Even in a less subjective domain, such as price level, a user's particular preference of upscale versus affordable, may affect the expectation of what is expensive versus not.

Preferences in particular can be inferred by user vote rankings, stated preferences, and history of visiting or ranking particular genre/location pairs.

Capturing these differences in preference and behavior is integral for informed universal ranks as well as personalized recommendations.

We loosely incorporate the individualized behavior preferences and biases into our model, arguing that weighting w/ trust scores alone will not ensure that user rating scores will approximate the globally biased distributions Z_l for a given location l. Two factors are to be considered

- If user voting preferences for genres in a given location, demonstrate different marginal probabilities for members of G than the globally biased distributions Z_l, then re-weighting user votes by preference is necessary.
- If general user voting behavior for a given genre and location $U_{g,l}$ does not reflect the underlying prior for the *true values* V, then the distribution of individual votes need to be transformed.

We consider $\hat{R}_{U_{L,G}} s \in \mathbb{R}$ as the theoretical continuous distribution of ratings provided by users defined by discrete values with individual bias applied for a given for locations and genre. In an ideal scenario with a set of trusted users $U_{G,L}$ for a specific genre and location, we subtract global bias with the inverse of h, and apply individual de-biasing function

$$b(h^{-1}(\hat{R}_{U_{L,G}}, L, G), L, G) \sim V_{G,L}$$

to approximate a distribution of votes that emulates the prior for the *true value* V. In practice, we consider b to be a scaling function, so that if users in a given location and genre demonstrate very tight distribution around a value, the function will map to a wider range of values.

Similarly, $a(u, g, l) \in \mathbb{R}^{[0,1]}$ maps an individual user preference for a particular genre in a given location, where 1 is viewed as complete preference and 0 is absolute dislike. The preferences across users for a particular location can be weighted to match the prior of a given $p(G = g)$.

Informed Rankings. To add finer granularity to the displayed ranking$\hat{r_e}$ is to weight users votes by the individual users preference for the given entity e genre/location pair. The intuition behind this suggestion is that a user with a known preference for a given genre $g \in G$ and $l \in L$ will have informed expertise on the given entity e. Applying preference scores can ensure that users with low preference for a particular genre are down-weighted and hence possible inconsistency (or down-weighted) scores are removed.

$$z_e \sim \frac{\sum_{u \in U_e} b(r_{u,e}, g_e, l_e) * a(u, g_e, l_e) * t_u}{\sum_{u \in U_e} a(u, g_e, l_e) * t_u}$$

In the same manner, personalization can conceivably be incorporated into the system. In this scenario, the displayed rankings would be customized to a *target* user's specified preferences.

In order to accomplish this task, votes can be weighted by a *similarity* score between a target user of the system and the user vote participants. In this context the function $a(u, g, l)$ can be replaced with $S(u, u')$, where S is a similarity metric and u' is the target user fro display.

$$\hat{z}_{u',e} \sim \frac{\sum_{u \in U_e} b(r_{u,e}, g_e, l_e) * S(u, u') * t_u}{\sum_{u \in U_e} S(u, u') * t_u}$$

A simple similarity metric could be the cosine-distance between tensors of users voting history or utilizing another collaborative filtering technique. Depending on the system, similarity can be defined by voting history, outside information, user specified preferences or user visit information. It is outside the scope of this work, but to ease in computation, users can be associated with predefined preference clusters, as such the choice of rating displayed is cached to a small number of pre-defined entities.

5 FUTURE DIRECTIONS

The concepts of trust, voting behavior and preference all have overlaps in interpretation and effective modelling. We separate explicitly in this paper on the user level and the global level, to dissect mechanisms in which informed ordinal rankings can be extracted.

For future work, we note that computing user preferences, trust, and voting bias on an individual user basis may not be computationally feasible or desirable. Instead of calculating individual voting bias, preference and trust, such values can be approximated via distance to pre-defined voting behavior and preference clusters.

An alternative consideration is personalizing rankings in a voting ordinal display. Instead of displaying rankings formed by weighting preference for a given entity genre/location, the individual votes can be weighted by the end users similarity to the voting users preference cluster.

REFERENCES

[1] Nicola Barbieri. 2011. Regularized Gibbs Sampling for User Profiling with Soft Constraints. *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011), 129–136.

[2] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. 2009. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *WWW*.

[3] Bee-Chung Chen, Anirban Dasgupta, Xuanhui Wang, and Jie Yang. 2012. Vote Calibration in Community Question-answering Systems. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 781–790. https://doi.org/10.1145/2348283.2348388

[4] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? *arXiv e-prints*, Article arXiv:1805.12002 (May 2018), arXiv:1805.12002 pages. arXiv:stat.ML/1805.12002

[5] Thomas Hofmann. 2003. Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. ACM, New York, NY, USA, 259–266. https://doi.org/10.1145/860435.860483

[6] Yehuda Koren and Joseph Sill. 2013. Collaborative Filtering on Ordinal User Feedback. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. AAAI Press, 3022–3026. http://dl.acm.org/citation.cfm?id=2540128.2540570

[7] Gary P. Latham. 2012. *Work Motivation: History, Theory, Research, and Practice*. Sage Publisher.

[8] R. Likert. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology* 140 (1932), 1–55.

[9] Benjamin M. Marlin. 2003. Modeling User Rating Profiles For Collaborative Filtering. In *NIPS*.

[10] David Owen. 2018. Customer Satisfaction at the Push of a Button: HappyOrNot terminals look simple, but the information they gather is revelatory. *The New Yorker* (Feb. 2018). https://www.newyorker.com/magazine/2018/02/05/customer-satisfaction-at-the-push-of-a-button

[11] Lahari Poddar, Wynne Hsu, and Mong-Li Lee. 2017. Quantifying Aspect Bias in Ordinal Ratings using a Bayesian Approach. In *IJCAI 2017*.

[12] Drazen Prelec, Hyunjune Seung, and John McCoy. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541 (01 2017), 532–535. https://doi.org/10.1038/nature21054

[13] Vikas C. Raykar and Shipeng Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13 (March 2012), 491–518.

[14] Lu Ren, Lan Du, Lawrence Carin, and David Dunson. 2011. Logistic Stick-Breaking Process. *J. Mach. Learn. Res.* 12 (Feb. 2011), 203–239. http://dl.acm.org/citation.cfm?id=1953048.1953055

[15] David H. Stern, Ralf Herbrich, and Thore Graepel. 2009. Matchbox: large scale online bayesian recommendations.. In *WWW*, Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl (Eds.). ACM, 111–120. http://dblp.uni-trier.de/db/conf/www/www2009.html#SternHG09

[16] Hao Wang and Martin Ester. 2014. A Sentiment-aligned Topic Model for Product Aspect Rating Prediction. In *EMNLP*.

[17] Pu Wang, Carlotta Domeniconi, and Kathryn Laskey. 2009. Latent Dirichlet Bayesian Co-Clustering. 522–537. https://doi.org/10.1007/978-3-642-04174-7_34

[18] Xiaochi Wei, Heyan Huang, Chin-Yew Lin, Xin Xin, Xianling Mao, and Shangguang Wang. 2015. Re-Ranking Voting-Based Answers by Discarding User Behavior Biases. In *IJCAI*.

[19] Qianli Xing, Yiqun Liu, Jian-Yun Nie, Min Zhang, Shaoping Ma, and Kuo Zhang. 2013. Incorporating user preferences into click models. In *CIKM*.

Learning to Predict Population-Level Label Distributions

Tong Liu
Rochester Institute of Technology
Rochester, New York
tl8313@rit.edu

Akash Venkatachalam
Rochester Institute of Technology
Rochester, New York
av2833@rit.edu

Pratik Sanjay Bongale
Gleason Corporation
Rochester, New York
pbongale@gleason.com

Christopher M. Homan
Rochester Institute of Technology
Rochester, New York
cmh@cs.rit.edu

ABSTRACT

Machine learning problems are often subjective or ambiguous. That is, humans solving the same problems might come to legitimate but completely different conclusions, based on their personal experiences and beliefs. In supervised learning, particularly when using crowdsourced training data, multiple annotations per data item are usually reduced to a single label representing ground truth. This hides a rich source of diversity and subjectivity of opinions about the labels. Label distribution learning associates for each data item a probability distribution over the labels for that item, thus can preserve the diversity that conventional learning hides or ignores. We introduce a strategy for learning label distributions with only five-to-ten labels per item by aggregating human-annotated labels over multiple, semantically related data items. Our results suggest that specific label aggregation methods can help provide reliable representative semantics at the population level.

CCS CONCEPTS

• **Theory of computation** → **Unsupervised learning and clustering**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*; • **Computing methodologies** → Classification and regression trees.

KEYWORDS

Subjectivity in crowdsourcing, label distribution learning, clustering and classification, social media.

ACM Reference Format:
Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to Predict Population-Level Label Distributions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317082

1 INTRODUCTION

The goal of many supervised learning problems is to map each given data item to a single (or set of, but in any case, deterministic) label(s) according to some standard of ground truth. However, many real-world problems—such as those related to color, pain, taste, level of danger, or qualitative analysis—have different answers depending on whom is asked, even when the domain of answers is fixed (i.e., *closed domain*) or more than one answer is allowed (i.e., *multilabel*). In such cases, a single (set of) label(s) does not meaningfully solve the problem, or may hide important dissenting beliefs or opinions. Yet the impact of AI agents that fail to recognize diversity in a representative fashion ranges from banal to harmful on a societal level.

For instance, in 2016 contestants from over 100 countries from around the world submitted images of themselves to Beauty.ai's website, and their proprietary deep learning agent, trained on publicly available facial images, chose winners in 44 different beauty pageant categories [28]. Yet the algorithm, perhaps due to biases in the trained data, showed strong signs of racial bias: 37 of the winners had distinctly European facial features [28]. Microsoft built a Twitter bot called Tay that was supposed to learn new language skills, but it had to be shut down soon after launch because it learned to deny the holocaust and demonize feminism [36]. ProPublica reported that Northpointe risk assessment software, used to by judges in Florida to help determine incarceration lengths, systematically assigned higher risk scores to black defendants than to white ones [31].

Label distribution learning (LDL) is a recent approach that replaces the goal of predicting, for each data item, a single (set of) label(s) with the more challenging and complex task of predicting a probability distribution (known as a *label distribution*) over the label choices [13]. A growing body of work has used this approach, e.g., to predict beauty in images [31] and rate movies [14], leading to more nuanced prediction results and insights. There is also evidence that, even in situations where ground truth exists but is difficult to obtain, predicting label distributions is more informative and accurate than aggregating the opinions of multiple labelers into a single (set of) discrete choice(s) [16].

A major resource bottleneck in LDL is the amount of human annotations needed, since for any large population of labelers, the number m of labels needed to estimate (i.e., taken as a sample of) the underlying population's true distribution of beliefs for even one data item is rather large, depending on the size of the label space and the degree of confidence needed. But the number of data items n needed for supervised learning normally runs into the thousands. Thus, taken independently, the total number $m \times n$ of human labels needed for training on label distributions grows quadratically in

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317082

two variables that are typically rather large, where their product can run into the millions or even billions.

Our main contribution is a new LDL strategy for reducing the total number of human labels needed per data item, by grouping together items determined to be semantically similar and pooling together their labels of all items in the each semantic class into a single label distribution which can be shared by all members of the class. Figure 1 illustrates this strategy.

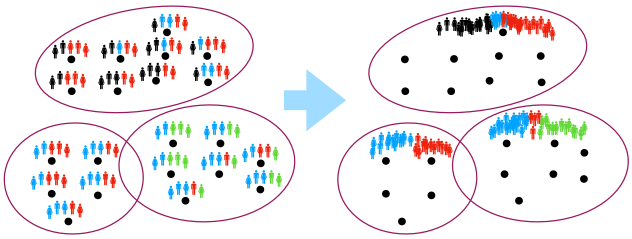


Figure 1: The main strategy this paper explores. The black dots represent data items. (Left:) Five labelers annotate each data item, where the color of the person indicates the label that person chose. If we view these five labels as a sample of the underlying population's beliefs, the sample size is probably too small for there to be much confidence in the sample. (Right:) We cluster together (indicated by the circles) semantic similar data items, and then pool together all the labels in each cluster into a single, larger sample which, according to our strategy, is a good representation of—and thus label distribution for—the population-level beliefs about each item in the cluster.

Specifically, we:

(1) Establish the premise for our proposed approach through a real-world example where there is substantial disagreement over the annotators' interpretations of 50 data items in a common social domain, but where the label distributions appear visually in histogram to cluster into a limited number of distinct classes.
(2) Introduce a novel family of algorithms for label distribution learning on as few as five-to-ten labels per data item that involves an unsupervised learning phase to yield hidden classes of semantically-related data items and assigns to each class an aggregated label distribution, followed by a supervised learning phase based on the labels the unsupervised phase produces.
(3) Show that, for larger label spaces, predictions based on unsupervised learning models that use our clustering strategy outperform those that do not, thus providing supervised learning validation for our approach.
(4) Perform our analysis on natural language data. We believe this is the first exploration of LDL on linguistic data from social media.

2 RELATED WORK

Probability and statistics play such an important role in machine learning that the field is sometimes regarded as a branch of applied statistics. However, in conventional machine learning, probabilistic models are typically used only for *performing* prediction; they are much less seldom the *objects* to be predicted.

Disagreement in human labeling tasks for supervised learning is widely studied as a common problem in its own right [9, 19, 25, 26, 33, 34]. Snow et al. [35], in a study on using multiple crowdsourced annotators to approximate the performance of experts, noted that individuals (including experts) tend to have personal biases (which later research [18] confirmed), but that multiple annotators may contribute to diversity, thus reducing individual annotator bias (see also [7, 11]). However, there is still an underlying assumption that a correct answer exists, even if it can never be directly confirmed.

Recent work has recognized the value of preserving subjectivity and ambiguity in data collection from human annotators. Aroyo and Welty show in a semantic parsing task that crowdworkers, when they agree with each other, can perform at a level comparable to domain experts, and when they disagree it is often for good reason, and in fact usually more desirable than collapsing to a single label [2]. Schaekermann et al. describe a framework for identifying unresolvable annotator disagreement [32].

Chen et al. [8] argue persuasively that to a wide spectrum of social scientists the volume of unstructured data available for qualitative analysis generated by social media is so great that automated methods like machine learning are needed to keep up. They also argue that preserving annotator disagreement is essential to applying qualitative methodologies like grounded theory at scale.

Zhang et al. [38] study the use of clustering to improve estimates of ground truth in crowdsourced labeling tasks, and show that the latent classes determined by their clustering models are, compared to plurality-based labels, better estimates of the semantics of the data items they study, thus providing support for this approach in the context of supervised learning (which they do not study). Clustering is a key subroutine in our learning algorithms, but here we study supervised learning tasks, where the goal is to predict the label distributions of unlabeled data (based on features separate from the labels themselves, in our case language as features), given a training set of human-labeled examples for learning. There are also key differences as to which clustering approaches we use and how we interpret the clusters they return.

Geng pioneered the systematic study of label distribution learning [13], where the objects to be predicted are probabilities of labels/classes. He and colleagues studied applications of LDL in many settings, some of which are related to predicting population-level distributions [14, 16, 31] while others are not [12, 15]. Several of these studies acknowledge the difficulty of obtaining valid label distributions that represent the underlying beliefs of human annotators; in fact, most of them were based on data and labels originally collected for the purpose of conventional (i.e., non-probabilistic labels) supervised learning problems. However, this line of research has thus far assumed that the label distributions obtained are equal to ground truth, i.e., without questioning the statistical validity of the data, even though the sample size of the labels for each item is small.

A number of research areas are related to LDL. In *multilabel learning* [39], each data item is associated with multiple labels. However, it does not typically distinguish between multiplicity due to disagreement (where different annotators might believe that only

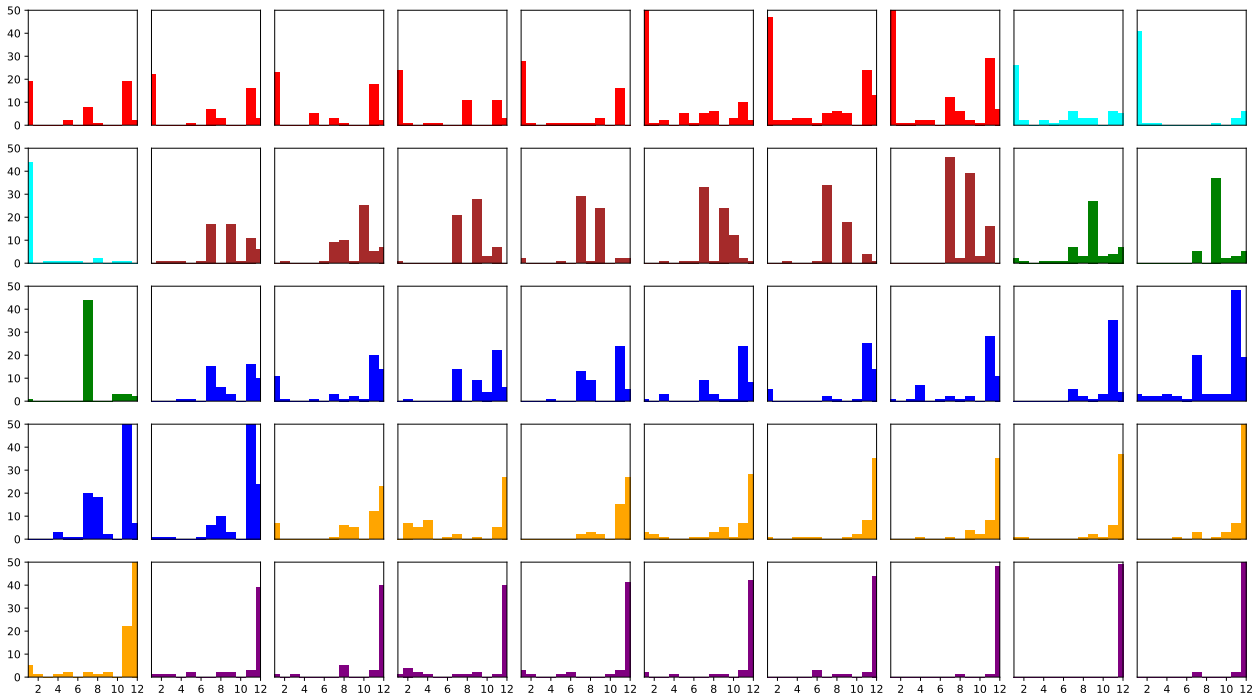


Figure 2: Histograms of the jobQ3MT+ label distributions. The X-axis ranges from 1 to 12, matching the Q3 choices in Figure 4. The Y-axis denotes the label counts. Tweets color: 1-8 red, 9-11 cyan, 12-18 brown, 19-21 green, 22-32 blue, 33-41 orange, and 42-50 purple.

.

one label is correct, but disagree on which one), ambiguity (where an annotator might believe multiple labels are valid), or uncertainty. Such distinctions may have significant social impacts, especially when disagreements fall along crucial demographic boundaries or indicate important but opposing perspectives that should just be preserved in the machine learning predictive models. Moreover, there are settings where label distributions are important but multi-label approaches do not naturally apply, such as when the prediction is ordinal (e.g., Likert-scaled) or real-valued. We are interested in capturing the diversity of beliefs across a population, where each member of the population may only associate a data item with one (set of) label(s), but different people may disagree on which ones.

One of the few areas where large samples of labeler beliefs or opinions are available are on commercial websites such as Amazon or Netflix. Recommender systems [3] seek to use such data to personalize user experiences like shopping, viewing, or playing by matching an individual's past behaviors and habits to similar individuals. Although modeling individual annotators is of interest to us (and a limitation of this paper is that we do not attempt to do so), in many data annotation settings there is, compared to recommender systems, little information available about the annotators to exploit. For instance, datasets labeled via crowdsourcing may not even have plenty of labels per annotator. Also, we are interested in predicting population-level beliefs about data items. Recommender systems, by contrast, predict which items an individual prefers.

There is research in the field of statistics on how to validly combine multiple samples from an underlying population [27]. These papers typically only consider standard statistical settings where a small number of samples are available and each is much larger than ours. We by contrast consider the problem of partitioning a massive number of very small samples into a much smaller number of clusters and then combining together all of the samples in each partition.

3 LABEL DISTRIBUTION LEARNING ON POPULATIONS

The *population label distribution learning problem* is to learn to predict the distribution of labels $\mathbf{y}$ among a population of annotators for each test set data item $\mathbf{x}$, given a collection of training data items $(\mathbf{x}_i)_{i\in\{1,\ldots,n\}}$ and a corresponding collection of label distribution *raw estimates* $(\hat{\mathbf{y}}_i)_{i\in\{1,\ldots,n\}}$, based on the normalized *empirical label distributions*, i.e., the distributions of the annotations received for each data item. Note that, here, we assume these distributions are multinomial samples of the underlying population of annotator's *true label distribution* $(\mathbf{y}_i)_{i\in\{1,\ldots,n\}}$, and that the each raw estimate was obtained by randomly choosing an annotator and then asking that annotator to choose a label, then repeating this process m times, where m is a parameter of the sampling process.

One example of a label set that supports this problem definition came from an effort to model Twitter discourse on life trajectories. When inspecting annotators' answers to a question which identifies

employment transition events, we observed that when there was disagreement it was often for good reason.

Figure 2 shows the label distributions over the the jobQ3MT+ label set (see more details in Section 4.1). These histograms of labels (one histogram per data item) appear to cluster into approximately eight categories, where the tweets in each seemed to be semantically related. Group 1 (red) distributions have most of their mass on *Getting hired/job seeking* and *None of the above, but job-related*, with tweets talking about plans to get a job (e.g., *really want a job*, *dont put that on ur resume for a minimum wage job*) or the process of getting a job. Group 2 (cyan) has almost all the mass exclusively on *Getting hired/job seeking* (e.g., *got the job*). Group 3 (brown) clusters around *Complaining about work* and *Going to work*, suggesting a topic about complaining about having to go to work. Group 4 (green) are a set of tweets complaining about work while at work. Groups 5 and 6 (blue and orange) have their peaks on *None of the above, but job-related* and *Not job-related*. Group 6 (where *Not job-related* was more frequent than *None of the above*) were mostly about road work. Group 7 seemed to contain cases where work was mentioned, but not central (e.g., *Today at work I learned about...*) or used "work" or "job" metaphorically, though there exist some clear *None of the above, but job-related* tweets, like *Perks of working overnight: donuts fresh out of the fryer*.

As to why such clustering happens, Zhang et al., on a different dataset, noticed similar clustering patterns [38]. We note that any k-choice annotation task effectively reduces the full breadth of interpretations encoded in each data item x to one of only k choices; We theorize that the act of annotation reduces not only the interpretive domain of the each data item, but also the social, experiential and cognitive factors, such as disparities in experience and knowledge, that drive annotator disagreement. Thus, the number p of *distinct* ground truth label distributions resulting from any annotation task are also limited, and the set of all annotations for any given data item is (assuming annotators are selected i.i.d. from the population of annotators) a sample from one of the p distinct ground truth distributions. We refer to this tentative explanation here as the *clustering theory*.

3.1 Overview of Learning Stages

Our approach to label distribution learning on populations consists of two stages. First, we use unsupervised learning to convert the raw label distribution estimates $(\hat{\mathbf{y}}_i)_{i\in\{1,\ldots,n\}}$, into *refined estimates* $(\hat{\mathbf{y}}'_\mathbf{i})_{i\in\{1,\ldots,n\}}$, by aggregating over semantic-related data items. Next, we perform supervised learning models on the refined label distributions with unstructured text features and conduct comparative experiments. We discuss each stage below.

3.2 Unsupervised Learning for Pooling Label Distributions

The unsupervised learning algorithms we consider here are consistent, to varying degrees, with the clustering theory. The (finite) multinomial mixture model **F**. This model most directly simulates the sampling process according to the cluster theory. It assumes that the empirical label distributions are generated by, (1) drawing a multinomial distribution π according to a Dirichlet prior over p elements (i.e., corresponding to the hypothesized number of true label distributions) $\pi \sim \mathrm{Dir}(p, \gamma = 75)$, where γ is the prior's (symmetric) hyperparameter (and higher numbers tend to produce lower entropy multinomials); (2) drawing multinomial distributions $\phi_1, \ldots \phi_p \sim \mathrm{Dir}(d, \gamma = 0.1)$; (3) for each data item, we (3a) choose $i \sim \pi$ and (3b) m labels according to ϕ_i. Thus, according to the clustering theory the most likely cluster distribution ϕ_j for each data item should be a good estimate of the true label distribution: $\phi_j \approx \mathbf{y}_i$. We use a variational Bayes algorithm[1] to learn the model.

Next come two variants of **F**. The Dirichlet process multinomial mixture model (**DP**) is a non-parametric version of **F**. Instead of choosing p multinomial models from a Dirichlet prior before generating the data, it starts with two multinomial models $\phi_1, \phi_2 \sim \mathrm{Dir}(d, 0.1)$. Then, for each new data item it draws from the current set of multinomial models in approximate proportion to the number of times each has been previously drawn OR draws a new multinomial model (with weight proportional to $\gamma = 50$). We use a variational Bayes algorithm to learn this model. The main purpose for including it here is to test whether in this setting non-parametric methods outperform parametric ones using standard model-selection criteria.

M is a multinomial mixture model without Dirichlet priors. This rather simple model can be learned using EM, however it lacks the regularization and adaptability that the Dirichlet priors provide. We expect this model to underperform the others.

In contrast to the previous models, we chose the Gaussian mixture model **G** as a weak alternate hypothesis of sorts. Rather than simulate the sampling process, as the multinomial distributions do, these distributions capture the variance in a population of samples. Additionally, it captures covariance between the labels; these should be close to zero in single label settings (or settings where the vast majority of annotators provide only one label per item). We use EM[2] to learn this model.

Finally, **L** is latent Dirichlet allocation[3] [4]. Though LDA is not a proper clustering model, we can obtain cluster-like latent classes from it. In terms of **F**, rather than choosing a single class selection distribution π for all data items, it chooses a new one π_i for each item i and for each label chooses a new distribution in $\{\phi_1, \ldots, \phi_p\}$ according to π_i. Thus, each instance of the labels for each item i from LDA represent a true mixture of all of the generating distributions, and is therefore not a proper clustering model (in contrast to the other models, where each instance of labels comes from one generating distribution only, although different instances may use different generators). Nonetheless, we can "assign" to i the most likely ϕ_j according to π_i.

3.3 Supervised Learning for Predicting Label Distributions

We train supervised-learning-based classifiers using refined label distributions obtained from the various unsupervised learning algorithms described above. We retain the most common 20,000 words in the test set and pad the sentence with up to 1,000 tokens in the text pre-processing step, then embed each word into a 100-dimension vector using the GloVe 2B-tweet corpus [29].

[1] Adapted from https://github.com/bnpy/bnpy.
[2] http://scikit-learn.org/stable/modules/mixture.html
[3] Based on https://radimrehurek.com/gensim/

We consider two neural network models. One ("**CNN**") is based on a 1D convolutional neural network, designed for sentence or tweet classification [21], with three max pool/convolution layers, followed by a dropout and a softmax layer. The other ("**LSTM**") is an encoder-decoder sequence-to-sequence model using recurrent neural networks . The encoder outputs a fixed-length encoding of the input text, and the decoder predicts the output sequence.

For both types of models, we use *softmax*: $\frac{exp(z_i)}{\Sigma_t exp(z_t)}$, to transform the output of the penultimate layer $\mathbf{z}$ into a probability distribution. We use *Kullback-Leibler (KL) divergence*, a standard measure of the difference between the "true" (in our case the refined estimate) probability distribution $\hat{\mathbf{y}}'$ and a predicted estimator $\tilde{y}$: $\mathrm{D}_{KL}(\hat{y}', \tilde{y}) = \sum_i P(\hat{y}' = i) \frac{\log P(\hat{y}'=i)}{\log P(\tilde{y}=i)}$, as the loss function for backpropagation, with the *Adam* optimizer [22]. We train with a batch size of 32 and 25 epochs.

4 EXPERIMENTS

Figure 3 summarizes our experiment framework which includes data and label collection, unsupervised and supervised modeling phrases and corresponding performance evaluations.

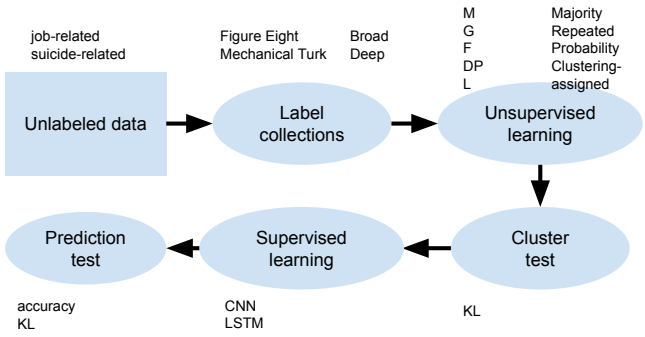


Figure 3: Our experimental workflow involves obtaining crowdsourced labels for raw data (yielding empirical label distributions for each data item), trying various unsupervised strategies for aggregating those labels, and finally testing how each approach affects the efficiency of supervised learning prediction. Note there are two testing phases: one for how well each aggregation strategy fits the data and one for supervised learning performance. We also list key terms, keywords, and abbreviations associated with each phase of the workflow.

4.1 Data and Labels

We consider two corpora: a set of 2,000 job-related tweets (mentioned in Section 3) and another set of 2,000 *suicide-related* tweets. Our institutional review board determined that our work did not fall under federal or institutional guidelines as human subjects type of research. Nonetheless, we took extra precautions to guarantee the privacy of the Twitter data: we replaced all mentions of usernames with "@SOMEONE" and URLs with "http://URL," and adhered to Twitter's developer policy[4]. Table 1 describes the basic properties of labels we collected for these two corpora.

Job-related. We introduced the job dataset in Section 3. It contains 2,000 tweets about work that were extracted by a publicly available library [24]. We asked five crowdworkers each from Figure Eight[5] (FE) and Amazon Mechanical Turk[6] (MT) to answer three questions about each tweet. Figure 4 shows the three questions we asked and their corresponding selections of labels. We denote these label sets *jobQ1/2/3*. To provide some insight into how performance might change with more labels from a more diverse population of labelers and labeling platforms, we first consider FE and MT as two separate label sets, then combine them into a single label set (denoted BOTH).

For each question, we then run experiments on two different train/dev/test splits. We first consider a 1000/500/500 split on each of the label sets: Q1, Q2, and Q3 (which we call the **Broad** split). Next, to get a more accurate ground-truth estimate for testing we randomly selected 50 tweets from our dataset and asked 50 additional MT crowdworkers to label them. We denote these label sets *jobQ1/2/3MT+* and create 1500/450/50 splits (called the **Deep** splits), where the training and development sets are from the BOTH label sets (minus the *jobQ1/Q2/Q3MT+* label set items) and the test sets are from *jobQ1/Q2/Q3MT+*, respectively.

Suicide-related. The *Suicide* tweet label set was obtained directly from [23]. It contains for each data item labels from five Figure Eight crowdworkers and up to two experts in suicide prevention. Each tweet was labeled as one of the following: Ⓐ *Suicidal thoughts*, Ⓑ *Supportive messages or helpful information*, Ⓒ *Reaction to suicide news/movie/music* and Ⓓ *Others*. We use a 1000/500/500 train/dev/test split.

4.2 Unsupervised Learning Experiments

4.2.1 Model Selection. For those clustering models requiring p as a hyperparameter, we test values for $p \in [d/2, 2d]$, where d is the number of label choices. We use the native likelihood function as our model selection criterion, because it is the native optimization goal of each unsupervised clustering algorithm. As the estimators for these models are stochastic and/or sensitive to initial conditions, for every model and every set of hyperparameters we ran 100 trials on the training/dev set and picked the model with the highest estimated likelihood. Table 2 shows the number of clusters selected on each of the two training splits on each label set and for **DP** the number of clusters the algorithm generated.

4.2.2 Evaluation. For the model M produced by each unsupervised learning algorithm and each data item i in the test set, we determine the most likely cluster j for i's empirical label distribution ϕ_j: $\arg\max_j P(\hat{y}_i \sim \phi_j \mid M)$. We then compute the KL divergence between the empirical label distribution $\hat{y}_i$ and the cluster distribution ϕ_j.

Table 3 shows that the multinomial mixture models (**M**/**F**/**DP**) generally outperformed **G**, as we expected. The crowdsourced sample sizes of 5–10 labels we used for each training item are typical

[4]https://developer.twitter.com/en/developer-terms/agreement-and-policy
[5]https://www.figure-eight.com/
[6]https://www.mturk.com/

Label Set	#Items	#Choices /item	#Workers	#Labels	Density	MVTD	RMSD
jobQ1FE	2,000	5	171	10,000	5.00	0.37	0.21
jobQ1MT	2,000	5	1,014	12,202	6.10	0.17	0.10
jobQ1BOTH	2,000	5	1,185	22,202	11.10	0.29	0.16
jobQ1MT+	50	5	249	2,969	59.38	0.43	0.22
jobQ2FE	2,000	5	171	10,000	5.00	0.28	0.16
jobQ2MT	2,000	5	1,014	12,202	6.10	0.15	0.09
jobQ2BOTH	2,000	5	1,185	22,202	11.10	0.23	0.13
jobQ2MT+	50	5	249	2,969	59.38	0.34	0.19
jobQ3FE	2,000	12	171	10,967	5.48	0.45	0.16
jobQ3MT	2,000	12	1,014	12,900	6.45	0.28	0.10
jobQ3BOTH	2,000	12	1,185	23,867	11.93	0.40	0.14
jobQ3MT+	50	12	249	3,196	63.92	0.41	0.14
Suicide	2,000	4	124	13,175	6.59	0.27	0.17

Table 1: Basic properties of our label sets. For the job-related data set with three questions *jobQ1/2/3*, *FE* and *MT* represent the labels from the platforms Figure Eight and Amazon Mechanical Turk respectively. *BOTH* combines both FE and MT labels. Density is the average number of labels per data item. MVTD (majority-voted-true-class deviation) and RMSD (root-mean-square deviation) are two proposed measures for estimating the variety and divergence of different label sets, motivated by the literature on scale and outlier description [20, 30, 37]. MVTD is the average deviation of the majority-voted label over all data items: MVTD = $1 - \sum_{i=1}^{n} \max_j\{\hat{y}_{ij}\}/n$. RMSD is the L2 deviation from the average label distribution: RMSD = $\sum_{i=1}^{n} \sqrt{(\hat{\mathbf{y}}_i - \overline{y})^T(\hat{\mathbf{y}}_i - \overline{y})}/n$, where $\overline{y}$ is the average label distribution over all data.

Q1. Which of the following items could best describe the point of view of job /employment-related information in the target tweet?

- ○ 1st person
- ○ 2nd person
- ○ 3rd person
- ○ Unclear
- ○ Not job-related

Q2. Which of the following items could best describe the employment status of the subject in the tweet?

- ○ Employed
- ○ Not Employed
- ○ Not in Labor Force
- ○ Unclear
- ○ Not job-related

Q3. Does the subject specifically mention any job/employment transition event in the tweet? (Choose all that apply)

- ☐1 Getting hired/job seeking
- ☐2 Getting Fired
- ☐3 Quitting a job
- ☐4 Losing job some other way
- ☐5 Getting promoted/raised
- ☐6 Getting cut in hours
- ☐7 Complaining about work
- ☐8 Offering support
- ☐9 Going to work
- ☐10 Coming home from work
- ☐11 None of the above, but job-related
- ☐12 Not job-related

Figure 4: The job-related annotation tasks contain these three questions and corresponding choices. The answers for Q3 are the columns in each of the histograms in Figure 2.

	Broad split					Deep split				
Dataset	M	G	L	F	DP	M	G	L	F	DP
jobQ1FE	10	4	9	3	4	11	11	9	3	4
jobQ1MT	11	4	11	8	10	2	2	11	9	11
jobQ1BOTH	11	2	2	6	8	2	2	11	7	8
jobQ2FE	11	3	10	3	4	11	2	10	3	4
jobQ2MT	2	4	11	7	9	2	2	11	7	10
jobQ2BOTH	2	2	11	5	7	2	2	8	5	7
jobQ3FE	19	5	18	6	7	19	10	19	7	7
jobQ3MT	5	5	14	17	20	5	19	15	17	26
jobQ3BOTH	5	15	18	13	16	5	17	11	17	17
Suicide	8	2	7	4	5	-	-	-	-	-

Table 2: The optimal label aggregation models on each label set using two splits (*Broad* and *Deep*) are achieved with the presented number of clusters (p).

of crowdsourced supervised learning label sets, and the differences between **G** and the other cluster models appear to be substantial at this scale. The success of **L** on a number of label sets surprised us, considering that we only use the mostly likely cluster for each data item which was trained on a mixture of clusters. Finally, **F** outperforms the other models on all of the sets having at least ten annotations per item, and shows the most improvement from the FE/MT (which had five annotations per item) to the BOTH (with ten annotations per item) label sets.

Table 3 also shows the average and standard deviation of the KL divergence scores on the four independent label sets (i.e., BOTH comprises FE and MT) jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide (highlighted in gray). These statistics indicate that **F** outperforms the other models across different thematic label sets in

KL	Broad split					Deep split				
	M	G	L	F	DP	M	G	L	F	DP
jobQ1FE	0.35	0.53	0.23	0.39	0.39	0.30	0.57	0.24	0.37	0.39
jobQ1MT	0.19	0.68	0.18	0.13	0.15	0.20	0.39	0.07	0.09	0.10
jobQ1BOTH	0.20	0.46	0.40	0.19	0.19	0.21	0.38	0.06	0.06	0.07
jobQ2FE	0.26	0.54	0.19	0.32	0.32	0.24	0.65	0.20	0.28	0.28
jobQ2MT	0.36	0.74	0.15	0.10	0.10	0.26	0.50	0.09	0.11	0.13
jobQ2BOTH	0.28	0.51	0.17	0.16	0.16	0.25	0.48	0.09	0.08	0.08
jobQ3FE	0.51	1.00	0.52	0.59	0.64	0.29	0.97	0.27	0.41	0.41
jobQ3MT	0.50	1.15	0.33	0.26	0.29	0.20	0.51	0.17	0.28	0.21
jobQ3BOTH	0.45	0.82	0.35	0.32	0.33	0.18	0.64	0.18	0.12	0.13
Suicide	0.22	0.57	0.20	0.22	0.22	-	-	-	-	-
Average	0.29	0.59	0.28	0.22	0.23	0.21	0.50	0.11	0.09	0.09
Std dev	0.10	0.14	0.10	0.06	0.06	0.03	0.11	0.05	0.02	0.03

Table 3: KL divergence based on the chosen label clustering models in Table 2. Average and standard deviation are based on the KL divergence scores of the gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide). The *lowest* KL is highlighted in yellow for each split.

its capability and stability, **DP** is second, and, as we expected, **G** comes last.

Readers may have notice in Figure 4 that Q3 differs Q1 and Q2 in allowing annotators to choose multiple labels. Theoretically, the ideal representation for the annotations (where each *annotation* is the set of labels provided by one annotator for one data item) of Q3 would be over the *power set* of possible labels. However, Table 4 shows that fewer than 10% of the annotations we received had selected more than one label. Thus, we simplify matters by representing these distributions as over just the base label set (i.e., *not* the power set), and treat multiple labels from the same annotator as if each came from its own, independent annotator (for example, an annotation with three labels provided is treated as three separate annotations.).

Label Set	Number of workers with labels they submitted				
	1	2	3	4	5+
jobQ3FE	10,000	722	176	53	16
jobQ3MT	12,202	628	58	11	1
jobQ3MT+	2,969	193	32	2	0

Table 4: Stats for workers annotating jobQ3 with numbers of labels they submitted individually.

4.3 Supervised Learning Experiments

We then trained the two supervised learning algorithms described in Section 3.3 on our training datasets' texts, using in turn each of the unsupervised learning methods described previously to provide *refined label distribution estimates* ($\hat{\mathbf{y}}'_i$) as the learning goal. We compared their performances to those of three common baseline strategies for resolving (or not) label disagreement.

- Majority (**Maj**) takes the final label to be $\hat{y}'_i = \underset{j \in \{1,\ldots,d\}}{\arg\max} \{\hat{y}_{ij}\}$.
- Repeated (**Rept**) duplicates each data instance once for every annotation it receives and pairs the replicated instance with that label.
- Probability (**Prob**) is the raw label distribution estimates ($\hat{\mathbf{y}}'_i$) = ($\hat{\mathbf{y}}_i$). (This is the baseline LDL approach.)

4.3.1 Evaluation. We measure the **KL divergence** between the classifier ($\tilde{\mathbf{y}}_i$) and cluster-or-baseline-method ($\hat{\mathbf{y}}'_i$) -based label distributions. (Note that Maj and Rept both associate each data item, by eliminating labels or creating copies of the data items, exactly one label. For the purpose of computing KL divergence we regard this as a distribution where the entire probability mass is on one label.) We also measure accuracy (**ACC**), i.e., the percentage of times $\arg\max_j \tilde{y}_{ij}$ matches $\arg\max_j \hat{y}'_{ij}$ in the test set. Accuracy is often used in nondistributional classification problems. We use it here to shed further light into the differences between distributional and nondistributional problems. In particular, we might expect that nondistributional models might outperform label distribution models with respect to accuracy, even as they underperform with respect to KL divergence.

4.3.2 Results. Tables 5-8 show the KL divergence and accuracy metrics for CNN/LSTM text classifiers built with different label aggregation strategies in two split modes (Broad split: Table 5 and 7, Deep split: Table 6 and 8) .

Starting with the KL divergence results, on the Broad split tests, CNNs trained and tested on **L** outperform other clustering and non-clustering approaches most of the time for both job and suicide discourse themes. For LSTMs, we can also observe that clustering approaches achieved better results more often on different label sets than non-clustering methods. Almost none of CNNs or LSTMs trained on any baseline label reduction strategy can compete.

By contrast, the results of the Deep split KL divergence tests (Table 6) are not as conclusive, and this could be due to there being fewer data items in the Deep split test set. But even so, clustering strategies again perform better in more cases than the baselines.

Tables 7 and 8 show that, for both the CNN and LSTM classifiers and both split modes, the highest accuracies often come from the clustering methods. They outperform non-clustering methods by more than 10% on average, which appears substantial. For those label sets whose accuracy based on clustering strategies do not rank 1st, non-clustering methods win only by a slim or zero margins.

Together, the results for different label sets and split modes reveal several interesting patterns. First, the cluster-based models tend to outperform the baseline methods in terms of either KL divergence or accuracy as reported. This supports the feasibility of our clustering strategy for label distribution learning on subjective problems with annotator disagreement. On the other hand, for conventional (i.e., non-distributional) classification problems, baseline methods can be sufficient, as shown in our experiment results. The advantages of clustering, in terms of KL divergence, is less stark in the Deep compared to the Broad splits, but clustering still seems to outperform baselines on the jobQ3 label set, which has the largest label space and is where pooling and other label conservation methods are most needed.

Broad - KL	CNN							
	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	2.98	0.79	0.91	0.12	0.74	0.47	0.18	0.19
jobQ1MT	2.03	0.80	0.72	0.65	1.05	0.52	1.02	1.00
jobQ1BOTH	2.38	0.45	0.48	0.36	0.38	0.27	0.40	0.38
jobQ2FE	2.29	0.91	0.79	0.21	0.78	0.13	0.31	0.28
jobQ2MT	2.10	0.80	0.78	0.81	0.98	0.67	1.04	0.96
jobQ2BOTH	2.12	0.49	0.47	0.48	0.48	0.37	0.51	0.52
jobQ3FE	4.20	1.66	1.14	0.31	0.68	0.66	0.42	0.36
jobQ3MT	3.18	2.24	1.05	1.04	1.32	0.54	1.12	1.12
jobQ3BOTH	3.38	1.40	0.77	0.62	0.49	0.62	0.71	0.70
Suicide	2.16	1.40	0.45	0.69	13.62	0.33	0.53	0.49
Average	2.51	0.94	0.54	0.54	3.74	0.40	0.54	0.52
Std dev	0.51	0.47	0.13	0.13	5.70	0.13	0.11	0.11
Broad - KL	**LSTM**							
	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	1.81	0.98	1.34	0.60	1.22	1.02	0.54	0.75
jobQ1MT	1.42	1.17	1.96	1.55	1.67	2.08	1.54	2.07
jobQ1BOTH	1.23	1.10	1.32	1.19	0.69	1.50	1.06	1.03
jobQ2FE	1.41	1.86	1.69	1.57	1.49	1.03	1.44	1.02
jobQ2MT	2.04	1.96	2.20	1.30	1.53	2.39	2.37	1.62
jobQ2BOTH	1.52	1.65	1.54	1.06	1.28	1.33	1.75	1.93
jobQ3FE	1.65	1.92	1.59	0.98	0.99	1.09	1.05	1.09
jobQ3MT	1.80	2.29	2.07	1.82	1.80	1.46	2.01	1.72
jobQ3BOTH	1.67	1.71	1.68	1.30	1.11	1.53	1.36	1.16
Suicide	1.50	1.27	1.34	1.32	16.40	1.28	0.90	1.07
Average	1.48	1.43	1.47	1.22	4.87	1.41	1.27	1.30
Std dev	0.16	0.26	0.15	0.10	6.66	0.11	0.32	0.37

Table 5: Kullback–Leibler divergence of the Broad split. Average and standard deviation are based on the gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide). The *lowest* KL is highlighted in yellow.

Deep - KL	CNN							
	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	3.09	0.77	0.90	0.13	0.69	0.39	0.09	0.16
jobQ1MT	2.94	0.47	0.54	0.64	1.08	0.47	1.22	1.05
jobQ1BOTH	2.90	0.34	0.24	0.39	0.43	0.38	0.33	0.35
jobQ2FE	3.07	0.57	0.65	0.18	0.56	0.49	0.21	0.31
jobQ2MT	1.90	0.50	0.58	0.77	0.68	0.76	0.74	1.07
jobQ2BOTH	2.90	0.27	0.28	0.52	0.37	0.35	0.50	0.58
jobQ3FE	3.71	1.45	1.00	0.34	0.63	0.65	0.53	0.43
jobQ3MT	3.95	1.98	0.77	1.13	1.21	1.20	1.26	1.24
jobQ3BOTH	3.33	1.13	0.63	0.76	0.67	0.49	0.71	0.73
Average	3.04	0.58	0.38	0.56	0.49	0.41	0.51	0.55
Std dev	0.20	0.39	0.18	0.15	0.13	0.06	0.16	0.16
Deep - KL	**LSTM**							
	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	1.16	1.06	1.01	0.92	0.95	0.80	1.16	1.17
jobQ1MT	0.93	0.85	1.09	1.38	2.04	1.54	1.61	2.52
jobQ1BOTH	1.30	0.91	1.20	1.20	1.49	1.13	0.82	1.09
jobQ2FE	1.59	1.11	1.58	1.23	1.38	0.96	1.17	1.09
jobQ2MT	1.16	1.00	0.76	1.69	1.50	1.51	1.42	2.47
jobQ2BOTH	1.02	1.10	1.17	0.97	1.19	1.63	1.69	1.26
jobQ3FE	1.86	1.90	1.67	0.76	1.02	1.73	0.93	1.20
jobQ3MT	1.38	1.39	1.69	2.22	1.65	2.25	2.13	1.59
jobQ3BOTH	1.64	1.49	1.46	1.69	0.99	1.94	1.37	1.44
Average	1.32	1.17	1.28	1.29	1.22	1.57	1.29	1.26
Std dev	0.25	0.24	0.13	0.30	0.21	0.33	0.36	0.14

Table 6: Kullback–Leibler divergence of the Deep split. Average and standard deviation are based on the gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, and jobQ3BOTH). The *lowest* KL is highlighted in yellow.

5 DISCUSSION

Our results provide evidence that clustering is a feasible strategy to improve performance of label distribution learning in certain settings, such as when each label distribution represents a population estimate based on a (micro) sample, and the data falls into a small number of semantic equivalence classes (relative to the learning task). Yet, why this is so is still not clear; our results shed little light on the validity of the clustering theory.

They also raise methodological issues. We expect that the methods introduced here for testing performance will provide helpful baselines for the development of newer methods tailored specifically toward settings where ground truth depends on a small number of samples per data item.

One methodological issue we grappled with was whether to measure the performances of the supervised models against the empirical label ($\hat{y}$) or the refined label ($\hat{y}'$) distributions. We felt that it was standard practice to test supervised learning on the patterns they are fed (i.e., the refined labels in our case), even though in our case the conventional machine learning algorithms are only the last half of a larger, novel supervised pipeline that happens to have an unsupervised front end, and which takes the empirical labels as input. We tried both approaches, but here, for space purposes and because we found our earlier results more interesting in this direction, we report on (and examined in much greater depth) only the predictions against $\hat{y}'$. The biggest worry in doing so is that, because pooling labels via a small number of clusters greatly reduces diversity in the label distributions, there is less likelihood of error, which would seem to make predictions easier when measure against $\hat{y}'$ than against the empirical distributions $\hat{y}$. (Certainly, if by chance there was only one cluster, the prediction task would be trivial.) While that may be so, in label distribution learning settings the more important question is, "What is the expected *degree* of error?" For this question, the relationship between label distribution diversity and performance is less clear.

Moreover, even when there is less diversity in the label distribution space, if the relationship between input data features and refined label distributions is less consistent than with the empirical label distributions (recall that the cluster algorithms were trained on the empirical label distribution, but the supervised models were trained on the natural language input data (using the output of the labels from clustering phase as the label distributions to learn), then the predictions could be less accurate.

We have been deliberately vague about what "population of labelers" means. This study was motivated by our work with microtask

	CNN							
Broad - ACC	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	0.73	0.53	0.72	0.78	0.95	0.58	0.64	0.58
jobQ1MT	0.80	0.72	0.79	0.56	0.67	0.76	0.54	0.56
jobQ1BOTH	0.82	0.64	0.81	0.57	0.76	0.76	0.65	0.64
jobQ2FE	0.73	0.63	0.79	0.71	0.62	0.94	0.59	0.64
jobQ2MT	0.73	0.68	0.73	0.48	0.55	0.71	0.53	0.52
jobQ2BOTH	0.76	0.65	0.76	0.63	0.58	0.71	0.54	0.56
jobQ3FE	0.36	0.31	0.41	0.47	0.32	0.45	0.42	0.46
jobQ3MT	0.53	0.45	0.51	0.26	0.28	0.49	0.28	0.28
jobQ3BOTH	0.48	0.42	0.53	0.31	0.62	0.46	0.25	0.21
Suicide	0.81	0.65	0.78	0.18	1.00	0.76	0.37	0.39
Average	0.72	0.59	0.72	0.42	0.74	0.67	0.45	0.45
Std dev	0.14	0.10	0.11	0.18	0.16	0.12	0.15	0.17
	LSTM							
Broad - ACC	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	0.85	0.74	0.86	0.89	0.98	0.76	0.83	0.79
jobQ1MT	0.87	0.83	0.86	0.79	0.80	0.86	0.78	0.77
jobQ1BOTH	0.87	0.79	0.88	0.79	0.88	0.84	0.79	0.82
jobQ2FE	0.87	0.80	0.88	0.79	0.79	0.96	0.83	0.80
jobQ2MT	0.84	0.82	0.85	0.76	0.77	0.85	0.76	0.75
jobQ2BOTH	0.84	0.81	0.86	0.81	0.80	0.83	0.76	0.76
jobQ3FE	0.67	0.65	0.67	0.72	0.67	0.68	0.67	0.71
jobQ3MT	0.69	0.69	0.71	0.63	0.65	0.64	0.65	0.63
jobQ3BOTH	0.70	0.68	0.67	0.63	0.77	0.70	0.64	0.63
Suicide	0.86	0.81	0.87	0.64	0.57	0.85	0.67	0.68
Average	0.82	0.77	0.82	0.72	0.76	0.81	0.72	0.72
Std dev	0.07	0.05	0.09	0.08	0.11	0.06	0.06	0.07

Table 7: Accuracy of the Broad split. Average and standard deviation are based on the gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, jobQ3BOTH and Suicide). The *highest* accuracy for each dataset is highlighted in yellow.

	CNN							
Deep - ACC	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	0.62	0.47	0.58	0.78	0.80	0.54	0.82	0.72
jobQ1MT	0.72	0.53	0.70	0.58	0.66	0.72	0.56	0.58
jobQ1BOTH	0.72	0.51	0.70	0.62	0.90	0.60	0.62	0.60
jobQ2FE	0.60	0.53	0.52	0.76	0.82	0.48	0.50	0.60
jobQ2MT	0.72	0.57	0.70	0.58	0.64	0.66	0.64	0.44
jobQ2BOTH	0.72	0.54	0.76	0.54	0.56	0.68	0.64	0.54
jobQ3FE	0.46	0.40	0.48	0.16	0.30	0.50	0.14	0.40
jobQ3MT	0.54	0.43	0.54	0.14	0.24	0.48	0.30	0.18
jobQ3BOTH	0.62	0.46	0.48	0.20	0.40	0.56	0.16	0.24
Average	0.69	0.50	0.65	0.45	0.62	0.61	0.47	0.46
Std dev	0.05	0.03	0.12	0.18	0.21	0.05	0.22	0.16
	LSTM							
Deep - ACC	Maj	Rept	Prob	M	G	L	F	DP
jobQ1FE	0.77	0.74	0.76	0.87	0.89	0.68	0.84	0.81
jobQ1MT	0.84	0.73	0.85	0.80	0.78	0.83	0.81	0.83
jobQ1BOTH	0.82	0.73	0.79	0.82	0.95	0.76	0.85	0.84
jobQ2FE	0.77	0.77	0.72	0.86	0.87	0.74	0.75	0.80
jobQ2MT	0.80	0.74	0.78	0.75	0.76	0.81	0.83	0.77
jobQ2BOTH	0.78	0.74	0.78	0.79	0.78	0.81	0.84	0.84
jobQ3FE	0.65	0.71	0.67	0.62	0.68	0.71	0.59	0.63
jobQ3MT	0.76	0.67	0.73	0.65	0.63	0.72	0.68	0.69
jobQ3BOTH	0.75	0.70	0.72	0.63	0.75	0.75	0.62	0.67
Average	0.78	0.72	0.76	0.75	0.83	0.77	0.77	0.78
Std dev	0.03	0.02	0.03	0.08	0.09	0.03	0.11	0.08

Table 8: Accuracy of the Deep split. Average and standard deviation are based on the gray-highlighted rows (jobQ1BOTH, jobQ2BOTH, and jobQ3BOTH). The *highest* accuracy is highlighted in yellow.

crowdsourcing sites like AMT and Figure Eight, in which case our labels can be taken as collection of (micro) samples of the population of workers on whichever sites are used for whatever interval of time the requested labeling task is posted. Studies exist on the demographics of these sites. Some sites (like Figure Eight in our study) provide some demographic information on the responders to each microtask request.

We have not yet modeled user behavior, though this is a well-established approach for aggregating labels from multiple annotators. We did, in fact, run experiments using Dawid and Skene's class annotator-based model [10], which is largely based on using behavior. However, as it is designed for conventional, non-distributional supervised learning and did not perform well, we did not report those results here. Another complication is that most of our annotators labeled only ten data items each, so we would be tempted to used clustering to group users in much the same way we used it here to group data items.

Another limitation was that we did not investigate in-depth the causes of inter-annotator disagreement, such as data encoding errors and communication ambiguities [1, 6, 40], lack of sufficient information [5, 6, 17], and unreliable annotators and their bias [17], nor did we attempt to resolve disagreement through follow-up discussions with the annotators, as is common in many grounded theory studies. We suspect in our experiment label sets, there exists some statistical correlations between the subjectivity and ambiguity and the degree of inter-rater disagreement across different questions. We hope to explore these directions in the future.

Another potential future direction could be to explore more highly structured label spaces, such as ordinal ones, or ones based on Bernoulli distributions or "single-peaked-ness" that are common in practice and sometimes yield to high-performance algorithms.

6 CONCLUSION

We study the important problem of predicting the distributions of population beliefs using both unsupervised and supervised learning methods. We test different strategies for clustering data items to obtain aggregated label distributions. We then build supervised CNN/LSTM classifiers using the predicted distributions and compared the performance with common baseline label reduction strategies. Our results from both unsupervised and supervised experiments show that it is feasible to predict probability distributions over labels at the population level. Clustering labels, in general, boosts the label distribution learning by aggregating data items with similar semantics and population-beliefs. We believe our study

is an pioneering exploration of disagreement on linguistic data from social media and further helps future intelligent agents understand the diversity of beliefs in society.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful feedback and comments. We thank Cyril Weerasooriya, Lingjia Deng, Chen Chen for helpful conversations.

REFERENCES

[1] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.
[2] Lora Aroyo and Chris Welty. 2014. The three sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34.
[3] Joeran Beel, Corinna Breitinger, Stefan Langer, Andreas Lommatzsch, and Bela Gipp. 2016. Towards reproducibility in recommender-systems research. *User modeling and user-adapted interaction* 26, 1 (2016), 69–101.
[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[5] Pavel Brazdil and Peter Clark. 1990. Learning from imperfect data. In *Machine Learning, Meta-Reasoning and Logics*. Springer, 207–232.
[6] Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research* 11 (1999), 131–167.
[7] Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 286–295. http://dl.acm.org/citation.cfm?id=1699510.1699548
[8] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (June 2018).
[9] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.
[10] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.
[11] Michael Denkowski and Alon Lavie. 2010. Exploring Normalization Techniques for Human Judgments of Machine Translation Adequacy Collected Using Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 57–61. http://dl.acm.org/citation.cfm?id=1866696.1866705
[12] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26, 6 (2017), 2825–2838.
[13] Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748.
[14] Xin Geng and Peng Hou. 2015. Pre-release Prediction of Crowd Opinion on Movies by Label Distribution Learning.. In *IJCAI*. 3511–3517.
[15] Xin Geng and Miaogen Ling. 2017. Soft Video Parsing by Label Distribution Learning.. In *AAAI*. 1331–1337.
[16] Xin Geng, Qin Wang, and Yu Xia. 2014. Facial age estimation by adaptive label distribution learning. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 4465–4470.
[17] Ray J Hickey. 1996. Noise modelling and evaluating learning from examples. *Artificial Intelligence* 82, 1 (1996), 157–179.
[18] Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, 107–117. http://www.aclweb.org/anthology/W14-3213
[19] Nicholas P Hughes, Stephen J Roberts, and Lionel Tarassenko. 2004. Semi-supervised learning of probabilistic models for ECG segmentation. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, Vol. 1. IEEE, 434–437.
[20] Rob J Hyndman and Anne B Koehler. 2006. Another look at measures of forecast accuracy. *International journal of forecasting* 22, 4 (2006), 679–688.
[21] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).
[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[23] Tong Liu, Qijin Cheng, Christopher Homan, and Vincent Silenzio. 2017. Learning from Various Labeling Strategies for Suicide-Related Messages on Social Media: An Experimental Study.. In *The workshop on Mining Online Health Reports of the 10th ACM Conference on Web Search and Data Mining*. Cambridge, UK. https://arxiv.org/pdf/1701.08796.pdf
[24] Tong Liu, Christopher M Homan, Cecilia Ovesdotter Alm, Ann Marie White, Megan C Lytle, and Henry A Kautz. 2016. Understanding Discourse on Work and Job-Related Well-Being in Public Social Media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1044–1053. https://www.aclweb.org/anthology/P/P16/P16-1099.pdf
[25] Andrea Malossini, Enrico Blanzieri, and Raymond T Ng. 2006. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics* 22, 17 (2006), 2114–2121.
[26] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* 19, 2 (1993), 313–330.
[27] Engineering, National Academies of Sciences, Medicine, et al. 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. National Academies Press.
[28] Jordan Pearson. 2016 (accessed November 11, 2018). *Why An AI-Judged Beauty Contest Picked Nearly All White Winners*. https://motherboard.vice.com/en_us/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners
[29] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162
[30] Robert Gilmore Pontius, Olufunmilayo Thontteh, and Hao Chen. 2008. Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics* 15, 2 (2008), 111–142.
[31] Yi Ren and Xin Geng. 2017. Sense beauty by label distribution learning. In *Proc, IJCAI*. 2648–2654.
[32] Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. 2016. Resolvable vs. irresolvable ambiguity: A new hybrid framework for dealing with uncertain ground truth. In *Proceedings of the 1st Workshop on Human-Centered Machine Learning at SIGCHI*.
[33] Padhraic Smyth. 1996. Bounds on the mean classification error rate of multiple experts. *Pattern Recognition Letters* 17, 12 (1996), 1253–1257.
[34] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems* 7 (1995), 1085–1092.
[35] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.
[36] Staff and agencies. 2016 (accessed November 11, 2018). *Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot*. https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot
[37] Cort J Willmott and Kenji Matsuura. 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science* 20, 1 (2006), 89–102.
[38] Jing Zhang, Victor S Sheng, Jian Wu, and Xindong Wu. 2016. Multi-Class Ground Truth Inference in Crowdsourcing with Clustering. *IEEE Trans. Knowl. Data Eng.* 28, 4 (2016), 1080–1085.
[39] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2014), 1819–1837.
[40] Xingquan Zhu and Xindong Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review* 22, 3 (2004), 177–210.

TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing

V. K. Chaithanya Manam
Purdue University
West Lafayette, Indiana, USA
vmanam@purdue.edu

Dwarakanath Jampani
Purdue University
West Lafayette, Indiana, USA
djampani@purdue.edu

Mariam Zaim
Purdue University
West Lafayette, Indiana, USA
mariam@zaeem.org

Meng-Han Wu
Purdue University
West Lafayette, Indiana, USA
wu784@purdue.edu

Alexander J. Quinn
Purdue University
West Lafayette, Indiana, USA
aq@purdue.edu

ABSTRACT

Developing instructions for microtask crowd workers requires time to ensure consistent interpretations by crowd workers. Even with substantial effort, workers may still misinterpret the instructions due to ambiguous language and structure in the task design. Prior work demonstrated methods for facilitating iterative improvement with help from the requester. However, any participation by the requester reduces the time saved by delegating the work—and hence the utility of using crowdsourcing. We present TaskMate, a system for facilitating worker-led refinement of task instructions with minimal involvement by the requester. Small teams of workers search for ambiguities and vote on the interpretation they believe the requester intended. This paper describes the workflow, our implementation, and our preliminary evaluation.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**; *Human computer interaction (HCI)*.

KEYWORDS

Crowdsourcing, task instructions, ambiguities, workflow

ACM Reference Format:
V. K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317081

INTRODUCTION

Crowdsourcing platforms offer the opportunity to complete digital work with quality levels which are only achievable with human power, but with the on-demand availability and scale of cloud computing. An oft-cited challenge for workers is ambiguous instructions, which may lead to seemingly wrong results for the requesters and frustration for workers [11]. Ultimately, this squanders the time and efforts of both the requester and worker.

Instructions for crowdsourcing tasks are generated in a process that is similar but typically less rigorous than the process used by expert annotators in behavioral sciences [15, 18]. In crowdsourcing, experts independently examine a sample of data, generate instructions specially around possible ambiguities identified in the sample data,and finally discuss and improve the instructions based on the feedback from others [12]. Writing comprehensive instructions that cover all the variations of the data in a dataset would require close examination of all the data which is typically impractical in the crowdsourcing settings.

Additionally, there exist several other methods to improve the quality of the work in crowdsourcing platforms while posting a task to multiple workers and later collecting their responses [9] via voting, ranking, or clustering. However, if the instructions are unclear, ambiguous, or do not provide enough information about the task, these existing mechanisms can lead to the rejection of conscientious efforts of workers, reducing efficiency of the system.

Ambiguities typically arise while to a novice requester, or a set of requesters try to transpose an idea following their own mental models into system-executable instructions, they may fail to explain their task clearly because of lack of time or skill. Classifying the ambiguities in a systematic way helps researchers study and resolve them efficiently. WingIt [16] presented a classification of ambiguities in the task into three major categories: input, process and output. Input ambiguity consists of form parameters that the requester supplies. Process ambiguity contains the procedural information about how to do the task, as well as any context provided by the requester. Output ambiguity contains the form fields, associated labels, and any text the worker is supposed to provide the requester.

Existing research on improving the quality of instructions in crowdsourcing can be classified into two major categories: reactive mechanisms [16] amd proactive mechanisms [3, 6, 7]. In a reactive mechanism, the requester will post all the tasks and when ever a worker identifies a problem, he/she will ask the requester and get clarification. All the clarifications provided by the requester are added to the original instructions while other workers are working on them. In reactive mechanism, the requester should be available and reply to the workers quickly while workers are working on the tasks until the task is completed by all workers. However, this might not be practical for busy requesters. In any proactive mechanism,

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317081

the requester posts a few instances of tasks and requests workers to provide feedback. Based on the feedback received from the workers, the requester improves the instructions. This is an iterative process and continues until requester gets correct results. Modify ing instructions through iterative processes based on the workers' feedback consumes lot of time for the requester. In this paper, we propose a workflow that delegates the task refinement to workers and thereby reducing the requester time.

This paper presents TaskMate, a workflow for engaging crowd-based workers to improve the quality of instructions for a given task. TaskMate allows workers to collaborate with each other to solve a task, therefore minimizing the effort and time that the requester needs to invest. To achieve this goal, we propose a workflow that divides the overall problem into small, manageable, and verifiable steps.

The key contributions of this paper are as follows:

- A novel workflow that will help busy requesters to create a high quality instructions in short time by recruiting a set of reliable crowd workers for a given task.
- TaskMate, a system that implements this workflow for information search tasks, the most common task type on Mechanical Turk [10].

RELATED WORK

Existing research on crowdsourcing emphasizes the impact of task instructions and design on the quality of the outcome. Gadiraju et al. identified the importance of task instructions in crowdsourcing and enhanced task clarity by modeling a solution with predictive features which would help enhance task clarity [5]. Wu et al. studied in detail how the quality of instructions will affect the quality of crowd work [19].

Task authoring interface enhancements

Fantasktic [8] is a system designed to help novice requesters design better tasks through three task design techniques: a guided task specification interface, a preview interface, and a worker tutorial. A guided task specification interface provides guidelines and recommendations to the requester when he/she is creating a task. A preview interface showcases the task as will be seen by a worker to the requester creating the task. A worker tutorial is generated automatically based on the sample answers provided by the requester. The quality of responses from workers showed significant improvement when the requester created a task following the guidelines of the guided task interface. However, task previews and worker tutorials did not have any impact on the workers' response. Guided task specification does not address the issue of ambiguities in the instruction. Moreover, it is not feasible to create guided interfaces for all of the different kinds of tasks on crowd platforms.

Worker-requester collaboration

Some recent approaches have demonstrated how workers and requesters can work together to efficiently improve an imperfect task design. Daemo [7] allows a requester to post few instances of the tasks to the workers and get their feedback. The requester can then improve the quality of his/her task based on workers' feedback. Accuracy of results were observed to be significantly higher for tasks that incorporated workers' feedback than for tasks that were originally posted.

Wingit is a reactive system in which a worker can edit the instructions or ask a question with a guessed answer to disambiguate their question when they encounter an ambiguity. The requester can then either trust the worker's guess and allow them to continue working on the task based on his/her guess—or the requester can review the worker's guess and reply with an acknowledgement containing the correct answer. In their evaluation, the workers demonstrated low proficiency identifying the ambiguities in the task instructions.

Structured labeling [13] is a tool that allows requesters to produce a clustered dataset where each cluster is labeled by a single person. Structured labeling enables requesters to independently learn about their task rather than through the worker-feedback.

Revolt [4] is a collaborative system developed to deal with ambiguous instructions in image-labeling tasks. Revolt produces structured labels created by the crowd. It allows multiple workers to label the images with specified instructions. In case of a conflict, workers relabel the images based on a description provided by other workers. The Revolt system is designed to improve the accuracy of results produced by workers only when there is a conflict. However, the Revolt system does not change the quality of instructions provided by the requester. Sprout [3] demonstrated a workflow that allows a requester to write minimum instructions for the task and then take the help of the crowd to create clear and detailed instructions. Taking feedback from workers and then improving the instructions for a task is extremely exhaustive in terms of both time and money for the requester.

With TaskMate, we show how workers acting without participation from the requester can identify and resolve ambiguities in the instructions, resulting in reduced response time while increasing overall operation efficiency.

Comparing TaskMate with WingIt

WingIt proposed Q&A and Edit methods for workers to resolve ambiguities in the instructions. The requester can either trust a single worker's judgment on the ambiguities or requester is available until the task is completed, to reply back to the workers' questions quickly when there is any ambiguity in the instructions. Trusting the judgement of a single worker may not always turn out to be the correct option. Additionally, the requester may not be available to reply to workers' questions in an on-demand basis. To overcome these problems, TaskMate employs multiple workers' judgments to resolve ambiguities in the instructions while the requester can choose to stay unavailable to the workers at any time during the process. By using WingIt, it is possible to post all the instances of the task and ambiguities are resolved while workers are working on the task. However, by using TaskMate, the requester need to post few instances of the task for improving the quality instructions. WingIt can resolve instance specific ambiguities but TaskMate may not.

TASKMATE

In this section, we present TaskMate, our system for improving the quality of instructions. The design decisions for TaskMate are based on previous work and the authors' hefty experience running crowdsourcing tasks.

TaskMate aims to identify ambiguities in the task, resolve them via majority voting, and generate clear instructions which clarifies all the ambiguities that are identified. TaskMate allows users to describe the task that they want to post on a crowdsourcing platform containing basic information. Users spending less than five minutes can receive high quality instructions within a short amount of time. We added an input decomposition at the beginning to identify all the list of possible ambiguities/problems following Soylent [2] and PlateMate [17]. The TaskMate framework consists of the following five stages: 1) Identify 2) Resolve 3) Merge 4) Verify 5) Select. The overall framework is composed of a multitude of iterative and parallel processes, as defined by Little et al. [14]. Figure 1 shows pseudo-code for the entire TaskMate workflow and Figure 2 explains the workflow with an example task.

Identify

The goal of this stage is to detect all the problems with the task instructions. Workers are presented with the original task instruction, asked to work on it, and then submit the results. When a worker is working on the task and finds any problem with the task, he/she will describe the problem in the form of question along with a set of possible answers. From the set of possible answers, the worker will choose one answer and complete the task. In the example task shown in Figure 2, it is not clear whether the requester is looking for top 3 computer science programs with in USA or Global or Europe. The worker will ask a question "Which ranking are you looking for?" with a list of possible answers (USA/Global/Europe). By the end of this stage, we will have a list of all problems with the given instructions along with a list of possible solutions for each problem.

Resolve

The goal of this stage is to guess correct solutions for each problem identified in the previous stage. Workers are presented with the original task ("List the top 3 Computer Science programs."), list of problems and the corresponding solutions to each problem. Workers were asked to guess what requester might be looking for from the list of possible solutions for each problem. Based on majority voting, we will get the correct solution for each problem. By the end of this stage, we will have a list of problems and solutions to each problem. In the example task, workers will vote for the two problems ("Which rankings are you looking for?", "Which level of college ranking are you looking at?"), based on majority voting, we will select the correct solution for each problem.

Merge

The goal of this stage is to incorporate all the problems identified in the original task and create a new task instruction. We can also have the task along with Q&A as given in Q&A method in WingIt [16]. However, the task instructions along with Q&A increases the length of the task and the workers need to spend more time reading and understanding the task. Previous studies [19] have shown that uptake rate will be less for lengthy tasks. So, to make tasks more concise or comprehensive, we ask the workers to merge the Q&A's in the task and create a new task instructions (e.g.,"List top 3 Computer Science Undergraduate programs in the world."). In order to get better instructions, we will ask more than one worker to combine original task and Q&A's. By the end of this stage, we will have a set of new task instructions ({"List top 3 Computer Science Undergraduate programs in the world.", "List the top 3 Computer Science business programs."}).

Verify

The goal of this stage is to verify whether all the ambiguities are correctly incorporated in to the new task instructions without changing the meaning of original task instruction. Workers are presented with original task, list of ambiguities along with solutions, and new task instructions that we got at the end of Merge stage. Workers were asked to verify either all the ambiguities are incorporated correctly or some of them are missing for each of the new task instructions. In the example task, workers will approve "List top 3 Computer Science Undergraduate programs in the world" and reject "List the top 3 Computer Science business programs". By the end of this stage, we will have a set of valid task instructions that incorporated all the problems in the original task instructions.

Select

The goal of this stage is to select a single task instruction from a set of task instructions that are clear and concise. Workers are presented with all the verified new task instructions and were asked to select one of them. Based on majority voting, TaskMate will finalize one task instruction ("List top 3 Computer Science undergraduate programs in the world."). This finalized instruction will be given to the requester, and he/she can use this instruction to post all the instances of the tasks on crowdsourcing platform.

System Usage

Our system is intended for requesters who aim to post a task on a crowdsourcing platform. With TaskMate, we hope to decrease the number of iterations that a requester has to go through in order to generate a clear set of instructions. Requester submits the task instructions to TaskMate, all the stages are triggered and posted to a crowdsourcing platform automatically. The requester can monitor the real-time progress of each of these stages. Once all five stages of TaskMate are complete, the 'new and improved' task is reported back to the user.

STUDY DESIGN

We conducted experiments on Amazon's Mechanical Turk (AMT) to investigate how TaskMate improves the quality of instructions provided through the system. Here, we describe the tasks, participant recruitment, and payment associated with our study.

To test TaskMate, we created tasks based on the taxonomy of ambiguities proposed in WingIt [16] and described them in Table 1. In our study, we planted ambiguities in the thirty tasks that comprise of all the different types of ambiguities. We are interested in studying the effect of each stage in TaskMate and we do not differentiate workers based on gender, geographic location, prior experience, and other personal characteristics. We randomly assigned three workers to work on each stage. Workers who worked on any stage are prevented from working on other stages for a given task.

We paid $0.40 for identify, $0.10 for resolve, $0.50 for merge, $0.10 for verify, and $0.15 for select per HIT. A total of 192 distinct

```
def improve_instructions(draft):
   # DETECT issues (using AMT)
   issues_qa = identify(draft, num_workers)
   # Each worker produces ≥ 0, Q → A+ pairs,
   # with ≥ 1 possible answers for each question.
   # issues_qa is like:
   #   [
   #     { "What size?" : ["big", "small"]},
   #     { "What color?": ["red", "green"],
   #       "How big?" : [">1 mile diameter", "small", "tiny"]}, ...
   #   ]

   # VOTE on resolution of issues (using AMT)
   changes = resolve(draft, issues_qa, num_workers)
   # changes is like:
   # { "What size?": "big", "What color?": "red", ... }

   # INTEGRATE changes
   # Several workers do this, resulting in multiple
   # candidate versions
   candidates = merge(draft, changes, num_workers)
   # Candidates is like:
   # set(["Find the big, green ball.",
   #     "Get the big green (?) ball(s).",
   #     "Get the bright, candy-colored ball that is small.",
   #     "Get skldfjsdflkjl.l;llllll;;", ...)

   # CULL wrong candidates
   correct_candidates = verify(draft, changes, candidates, num_workers)
   # correct_candidates is like:
   # set(["Find the big, green ball.",
   #     "Get the big green (?) ball(s).",
   #     "Get the bright, candy-colored ball that is not small.", ...)

   # SELECT best candidate
   winner = select(draft, changes, correct_candidates, num_workers)

   return winner
```

Figure 1: Pseudocode for the TaskMate workflow.

workers have participated in our study. Our experimental tasks, problems associated with each task, and the improved task are given in Table 2.

DISCUSSION & FUTURE WORK

Our study shows that workers are not able to detect ambiguities in the task significantly. This can be improved by training workers on identifying each of these ambiguities. In TaskMate, we observed that the task completion time is in the order of minutes. This can be due to the worker-demographics, worker-availability, the relative attractiveness of work, and the amount we pay. In order to reduce the completion time, we can either increase the amount per task or use retainer model [1]. We have not used any Natural Language Processing (NLP) techniques to validate the task provided by the

Input	Process	Output
Entity (IE)	Steps (PS)	Entity (OE)
Syntax (IS)	Words (PO)	Exception (OX)
Wrong (IW)	Wrong (PW)	Units (OU)
Units (IU)		Format (OF)
		Precision (OP)

Table 1: Classification of ambiguity in task instructions [16].

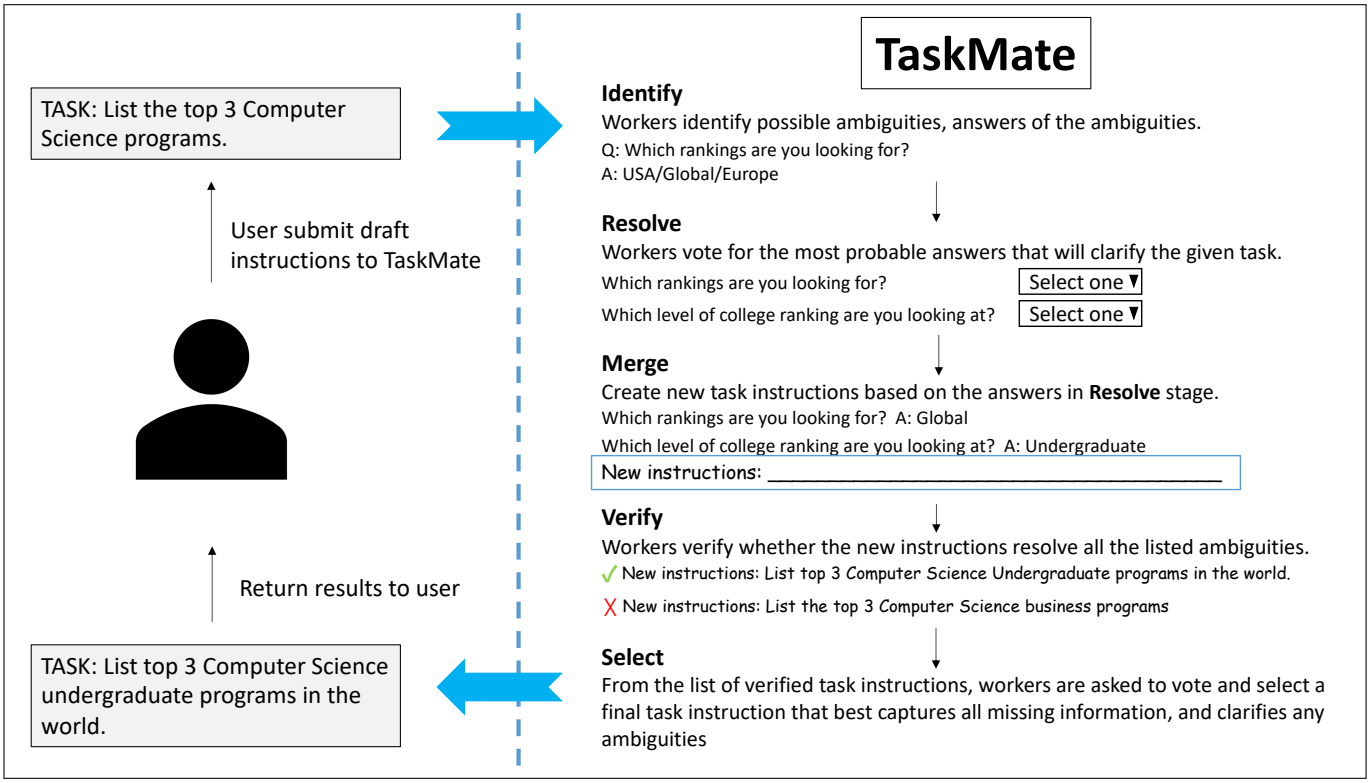


Figure 2: TaskMate workflow with an example task

requester. For future work, we could use NLP initially to identify some of the ambiguities by using word sense disambiguation. Furthermore, we can correct grammatical and syntactical errors in the original instructions automatically.

In the current system, workers need to identify all the ambiguities in the first stage itself. In the future, we can extend TaskMate as a cyclic process by asking the worker at each stage where any new ambiguities are found or not. If any new ambiguities are found, we can move back to the first stage and continue until all ambiguities are resolved. We've observed that workers are able to identify some of the problems in the task and improve the instructions based on the identified problems. We've also found that workers were able to identity ambiguities in the task and the list of possible solutions but they were not able to find the correct one that the requester wants in resolve stage. So, in the future, we can use the requester in resolve stage in order to achieve better results.

CONCLUSION

In this paper, we present TaskMate: a mechanism to improve the quality of instructions in crowdsourcing. TaskMate consists of five stages: Identify, Resolve, Merge, Verify, and Select. In the Identify stage, each worker will identify the problems with the task instructions and a set of possible solutions by working on the task. In Resolve stage, all the ambiguities identified in the previous stage are resolved by choosing the best possible solution to the resolve the ambiguity. In Merge stage, workers will write a new task by merging the original instruction along with the list of ambiguities and clarifications. In Verify stage, workers will verify whether the new merged task instruction has incorporated all the ambiguities associated with the set of instructions generated from previous stages as output of the original instruction. In the final stage, Select, workers will vote for the best instruction which does not contain any form of vagueness or lack of clarity. Our results show that workers were able to come up with instructions that are clear and precise.

REFERENCES

[1] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *UIST '11: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 33–42.

[2] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *UIST '10: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 313–322.

[3] Jonathan Bragg and Daniel S Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *UIST '18: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, ACM, New York, NY, USA, 165–176.

[4] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *CHI '17: Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2334–2346.

[5] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *HT '17: Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, 5–14.

[6] Snehalkumar Gaikwad, Nalin Chhibber, Vibhor Sehgal, Alipta Ballav, Catherine Mullings, Ahmed Nasser, Angela Richmond-Fuller, Aaron Gilbee, Dilrukshi

Gamage, Mark Whiting, et al. 2017. Prototype Tasks: Improving Crowdsourcing Results through Rapid, Iterative Task Design. *arXiv preprint* arXiv:1707.05645 (2017).

[7] Snehalkumar (Neil) S. Gaikwad, Mark E. Whiting, et al. 2017. The Daemo Crowdsourcing Marketplace. In *CSCW '17: Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. ACM, New York, NY, USA, 1–4.

[8] Philipp Gutheim and Bjorn Hartmann. 2012. *Fantasktic: Improving Quality of Results for Novice Crowdsourcing Users*. Master's thesis. EECS Department, University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-112.html [Online; accessed September 1, 2018].

[9] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *HCOMP '10: Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, New York, NY, USA, 64–67.

[10] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017), 829–840.

[11] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd Work. In *CSCW '13: Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, 1301–1318.

[12] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *CHI '14: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3075–3084.

[13] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. In *CHI '14: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 3075–3084.

[14] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 68–76.

[15] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. *CAM Journal* 10, 2 (1998), 31–36.

[16] VK Chaithanya Manam and Alexander J Quinn. 2018. WingIt: Efficient Refinement of Unclear Task Instructions.. In *HCOMP '18: Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing*. AAAI Press, Palo Alto, CA, USA, 108–116.

[17] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. 2011. Platemate: Crowdsourcing Nutritional Analysis from Food Photographs. In *UIST '11: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 1–12.

[18] Cynthia Weston, Terry Gandell, Jacinthe Beauchamp, Lynn McAlpine, Carol Wiseman, and Cathy Beauchamp. 2001. Analyzing interview data: The development and evolution of a coding system. *Qualitative sociology* 24, 3 (2001), 381–400.

[19] Meng-Han Wu and Alexander J. Quinn. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. In *HCOMP '17: Proceedings of the 5th AAAI Conference on Human Computation and Crowdsourcing*. AAAI Press, Palo Alto, CA, USA, 206–215.

#	Original Task	Instance	Problem and solution	Improved Task (TaskMate)
1	Find the URL for the following professor in Electrical and Computer Engineering at Purdue	Raghunathan	Q: Which one? (IE) A: Most senior professor; prefer Computer Engineering: Vijay Raghunathan Q: Which URL? (OE) A: Most specific home page	Find the personal URL for Vijay Raghunathan, a professor in Electrical and Computer Engineering at Purdue, on the universities information page.
2	Find the address of a movie theater close to the following airport in the city below	Chicago	Q: Which one? (IE) A: Largest in the area Q: How close? (OP) A: Closest based on driving time.	Find the address of a closest movie theater to the airport Chicago O'Hare.
3	When was the following movie released?	Harry Potter	Q: Which one in the series? (IE) A: Most recent Q: Date or just year? (OF) A: Date	What year was Harry Potter and The Sorcerer's Stone released in movie theaters in the United States?
4	Find the weight of the any smartphone with the following brand that can be purchased on BestBuy.com for $30000 to $49900.	Moto G	Q: Which model? (IE) A: Most recent Q: Do you mean "$300.00 to $499.00"? (PS) A: Yes. Q: Weight in g or oz? (OU) A: g	Go to BestBuy.com and do a search for Moto G brand smartphone (cell phone only) from the current year that is priced between $300-$499. From the detail listing page, copy the weight that is given.
5	Find the cost of attending this university for undergraduates.	MIT	Q: domestic or out-of-state? (OE) A: domestic Q: Total or individual components (tuition, fees, books, etc.)? (OP) A: Sum of tuition and fees only	Find the cost, including the tuition and fees, of attending the School of Architecture and Planning at MIT without a full scholarship or in-state residency discounts while living off campus.
6	Find the i-10 index and h-index for the following professors. Go to Microsoft Scholar and search for the person's name.	Michael Bernstein (Stanford)	Q: What are i-10 and h-index? (PO) A: Use Google Scholar Q: All or last 5 years? (OE) A: All	The i-10 index refers to the number of papers with 10 or more citations. The h-index is an author-level metric that attempts to measure both the productivity and citation impact of the publications of a scientist or scholar. Search Google Scholar for the i10 index and h-index since 2014 for the following professors: Michael Bernstein (Stanford).
7	Use search engine to find the home page URL for the following organization	WMT	Q: Are these stock market names of the companies Walmart (WMT)? (IE) A: Yes	Use search engine Google to find the home page URL for the following organization: Walmart (WMT).
8	Go to the Google Store (https://store.google) and search for the device below. Enter the price of the least expensive option.	Samsung Pixel	Q: Did you mean "Google Pixel"? (IW) A: Yes. Q: What model? (IE) A: Google Pixel 3 Q: Do you mean "store.google.com"? (PW) A: Yes.	Please, enter the Google Store website (https://store.google.com) and look for the price of the following device: Google Pixel 3. Write down the cheapest option you can find.

#	Original Task	Instance	Problem and solution	Improved task (TaskMate)
9	Find the maximum extendable ROM card size for the following devices	Moto G4 Plus	Q: Do you mean extendable memory card? (IW) A: Yes Q: GB, MB, inches, cm, mm? (OU) A: GB	Find the maximum extendable memory card size for the Moto G4 Plus.
10	Find a city that has an average temperture between 20 to 30 during May in the following state	Indiana	Q: Do you mean temperature? (IS) A: Yes Q: Celsius or Fahrenheit? (IU) A: Fahrenheit Q: What if no city in that state fits the criteria? (OX) A: Enter "none"	Please search in google and find and list all the cities in Indiana state, USA where More then 5000 residents reside and average temperature between 20 to 30 during May. Temperature must be in Fahrenheit.
11	Search for flight information from Chicgo to the following city in December with maximum price 700	Beijing, China	Q: Q: Do you mean Chicago? (IS) A: Yes Q: Cost in US dollars or Euros (IU) A: US dollars	Use Expedia to find the least expensive round-trip flight from Chicago, IL to Beijing, China during December. Include the name of the airline and the dates for each flight (maximum price $700 USD)
12	Find the URL for graduate program admission process in the following universities	MIT	Q: Where can I find it? What should I search for? (PS) A: Search in the admission page	Find the URL for the graduate Political Science admissions page at MIT for the year 2019.
13	Find three mobile phones from below manufacture that cost less than $300	Apple	Q: Where can I find it? What should I search for? (PS) A: Search in amazon or any shopping website	In US dollars, find 3 mobile phones (old or new) from Apple that cost less than $300.
14	Find the impact factor for the following journals	Transaction on Human Computation	Q: What is impact factor? (PO) A: Search online and find out	Find the impact factor of the academic journal "Transaction on Human Computation" on students, based on the number of times the academic journal has been cited, using journal citation reports, given that an average article is cited 100 times, as measured by Thomson Reuters.
15	Go to amazon.com and search for the best price of the following car, base model, year 2017	Honda CRV	Q: Old or New? (IE) A: New. Q: Do you mean cars.com or some car website instead of amazon.com? (PW) A: Yes	Go to Amazon.com and search for a GPS system for a 2017 Honda CRV. Find the best price.
16	Go to cars.com and find the best price for the following mobile phones	iPhone 6s	Q: Old or New? (IE) A: New. Q: Do you mean Amazon.com or some shopping website instead of Cars.com? (PW) A: Yes	Go to the Apple website and find the best price for the iPhone 6s.

#	Original Task	Instance	Problem and solution	Improved task (TaskMate)
17	Go to bestbuy.com and then search for the following key words. Send the top 3 links that you get	Buy a car	Q: Do you mean google.com instead of bestbuy.com? (PW) A: Yes	Go to bestbuy.com. If it asks for a country, use USA. Search for the following keyword: Buy a car. Enter the first 3 links that you get.
18	Find the URL and the fee for undergraduate engineering program at the following Universities	University of Texas, Austin	Q: Is it for domestic or international? (OE) A: International Q: Is it for Semester or Yearly or 4 years fee? (OE) A: Yearly Q: Total or individual components (tuition, fees, books, etc.)? (OP) A: tuition+fees only	Please find the URL and the total tuition cost for the all undergraduate engineering degree programs at the college of engineering at the University of Texas' Austin location only.
19	Find the city fuel efficiency of the following Indian Car	Mahindra XUV	Q: Manual or Automatic? (IE) A: Manual Q: Is mileage in MPG or Kilometers per liter? (OU) A: Kilometers per liter	Find the city fuel efficiency in kmpl of the following Indian Car: Mahindra XUV (petrol variant).
20	Find the average temperature of the following city in Decmber	Chicago, USA	Q: Do you mean December? (IS) A: Yes Q: Is temperature in Celsius or Fahrenheit. (OU) A: Fahrenheit	Find the average high and low temperatures in Chicago, USA during the month of December. Also provide precipitation averages if listed.
21	Find the fuel tank capacity of the following vehicle	Chevrolet Cruze	Q: Gallon or liters? (OU) A: Gallons	Find the fuel tank capacity in gallons of a Chevrolet Cruze produced after 2015 with a link to where you found the information.
22	Find the date when the 1st model of following phone was launched	Google Nexus	Q: Date format? (OF) A: MM/DD/YYYY	Find the date when the 1st model of the Nexus One was first available for purchase.
23	Find the launch date for the next model of the following mobile phone	Nexus Pixel	Q: What if new phone is not announced? (OX) A: Write ""Not announced"" Q: Date format? (OF) A: MM/DD/YYYY	Find the launch date for Google Pixel the next model.
24	Find the recipe of the following food	Fried Rice	Q: Which fried Rice? (IE) A: Chicken Fried Rice Q: URL or entire recipe? (IW) A: URL Q: How do i find the recipe? Do I have to search online or ask someone and enter it here (PS) A: Search online	Use google and find one recipe for Vegetable Fried Rice using boiled white rice (It should only take 10 minutes to cook) and that can be used in a fast food restaurant business.
25	Find three URLs for the best used cars that are priced maximum 7000 dollars with in 100 miles from ORD Chicago	Honda Accord 2007	Q: Which model? Base model or high end. (IE) A: Base model Q: What does ""best"" mean? (PO) A: Lowest Price	Find three URLs for the best used car below that are priced maximum 7000 dollars that allows loan payment within 100 miles from ORD Chicago: Honda Accord 2007.

#	Original Task	Instance	Problem and solution	Improved task (TaskMate)
26	Find the Official twitter account for the following celebrities	Kristen Stewart	Q: What if no account matches given criteria? (OX) A: enter "none"	Find the official, verified twitter account for Kristen Stewart, who is a famous worldwide celebrity. Look for the verified check mark that lets you know you found the official account for Kristen Stewart. When you have found the official account, copy and paste the URL of the twitter account into the field below.
27	Find the price of the following car in Chicago	Honda Accrd	Q: Coupe or Sedan? (IE) A: Sedan Q: New or Used? (IE) A: New Q: How to find the price? Should I search online or call the car company? (PS) A: Search online Q: Do you mean Honda Accord? (IS) A: Yes	Find the average price of the car below: New 2019 Honda Accord, Chicago IL.
28	Find the lowest price for the device in the US or Canada.	iPhone 7TT	Q: Did you mean "iPhone 7 Plus"? (IW) A: Yes. Q: How can I search by country? (PS) A: Search amazon.com and amazon.ca. Then take minimum of the two. Q: USD or CAN? (OU) A: USD	Find the lowest price for the new device iPhone 7 plus in the US or Canada the payment method can be made through debit.
29	Find names and URLs of the professors who's area of research is HCI in the following Universities	CMU	Q: How do I find the names? Via online search or contacting someone in the university? (PS) A: Search online Q: Do you need one or more than one Prof. name? (OE) A: only one	Look for the professor personal website URLs in charge of research about Human-Computer interaction in the Carnegie Mellon University. One professor per task.
30	Find the store name and the address near 465 Northwestern Ave, West Lafayette, IN 47907 where we can buy the following plants	Spider Plant	Q: Near means ? Is it 3 miles or 30 miles? (IE) A: Within 10 miles radius Q: What if no account that matches given criteria? (OX) A: enter "none"	Find the closest store with plants to 465 Northwestern Ave, West Lafayette, IN 47907.

Table 2: Experiment data

Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format

Mike Schaekermann
University of Waterloo
Waterloo, Canada
mschaeke@uwaterloo.ca

Graeme Beaton
University of Waterloo
Waterloo, Canada
graeme.beaton@edu.uwaterloo.ca

Minahz Habib
University of Toronto
Toronto, Canada
minahz.habib@mail.utoronto.ca

Andrew Lim
University of Toronto
Toronto, Canada
andrew.lim@utoronto.ca

Kate Larson
University of Waterloo
Waterloo, Canada
kate.larson@uwaterloo.ca

Edith Law
University of Waterloo
Waterloo, Canada
edith.law@uwaterloo.ca

ABSTRACT

Group-based discussion among human graders can be a useful tool to capture sources of disagreement in ambiguous classification tasks and to adjudicate any resolvable disagreements. Existing workflows for panel-based adjudication, however, capture graders' arguments and rationales in a free-form, unstructured format, limiting the potential for automatic analysis of the discussion contents. We designed and implemented a structured adjudication system that collects graders' arguments in a machine-readable format without limiting graders' abilities to provide free-form justifications for their classification decisions. Our system enables graders to cite instructions from a set of labeling guidelines, specified in the form of discrete classification rules and conditions that need to be met in order for each rule to be applicable. In the present work, we outline the process of designing and implementing this adjudication system, and report preliminary findings from deploying our system in the context of medical time series analysis for sleep stage classification.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**; **Human computer interaction (HCI)**; *Collaborative interaction*; *Empirical studies in collaborative and social computing*.

KEYWORDS

Ambiguity; Disagreement; Adjudication; Medical time series

ACM Reference Format:
Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317085

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317085

1 INTRODUCTION

A common requirement in supervised machine learning is that objects can be *unambiguously* classified into categories. In practice, however, there exist many classification tasks that are inherently ambiguous and the reasons why domain experts may be in disagreement over the correct way to classify an object may vary from task to task and from data object to data object. Several researchers have recognized this problem and come up with different solutions to handle it. One main distinction between these different works can be made around the question of whether expert disagreement is a problem to be resolved or whether disagreement is treated as a signal that is leveraged in some useful way. Our work is situated along the latter line of research. In particular, we propose that a key component to trusted and explainable artificial intelligence (AI) systems is to understand and capture the logical arguments and the various pieces of evidence that lead to divergent interpretations among experts. The overall goal is to get one step closer to endowing AI systems with the ability to provide argument-based explanations about (potentially ambiguous) classification decisions to their end users. Extending prior work on the design of systems for real-time group deliberation among remote human annotators [20, 22] and observational studies of in-person adjudication among expert annotators [21], in this work, we propose a general approach for capturing experts' rationale for individual classification decisions in a structured, guideline-centric format–with the goal of capturing sources of ambiguity and the content of evidence-driven adjudication discussions in a machine-readable format. The remainder of this paper covers related work, briefly introduces the reader to the application domain of sleep stage classification, details the proposed solution and preliminary findings from pilot experiments, and concludes with a discussion of use cases for our approach.

2 RELATED WORK

2.1 Ambiguity and Inter-rater Disagreement

Ambiguity is an issue of central importance in the field of epistemology, where an openness to multiple interpretations complicates the justification of knowledge. In practice, ambiguity gives rise to inter-rater disagreement in expert domains when there is lack of consensus on a single interpretation of a subjective case. Both ambiguity and expert disagreement have received extensive coverage in the epistemological literature [2, 7, 13, 24]. In a recent discussion of

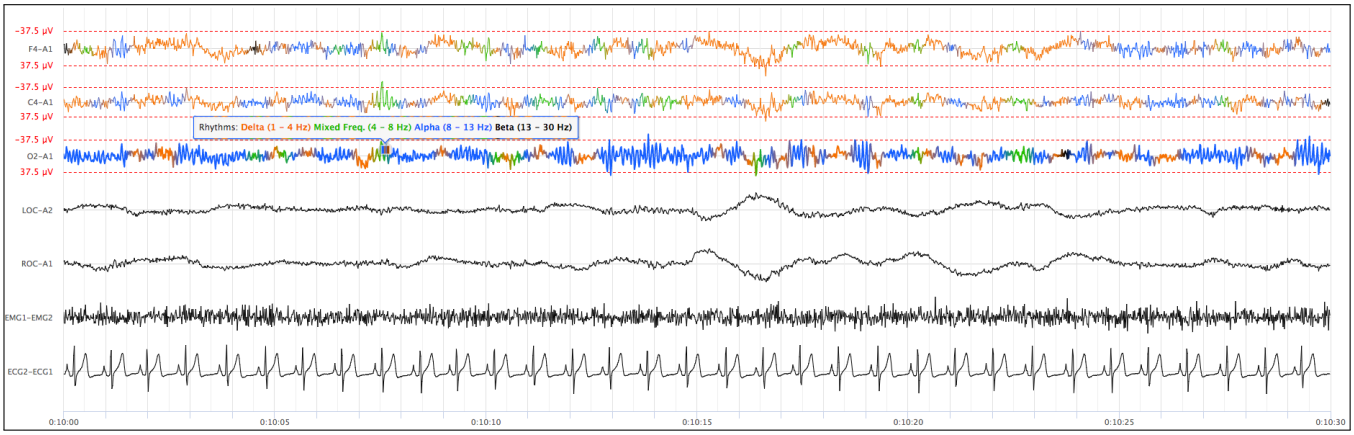


Figure 1: Visualization of one 30-second epoch of biosignal data to be scored into one of five stages of sleep.

the issue, Garbayo (2014) distinguished legitimate disagreement—where experts disagree despite access to the same evidence— from verbal disagreement, where experts misunderstand each other due to differences in terminology or semantics [13]. In an earlier theoretical account, Mumpower and Stewart (1996) delineate three forms of expert disagreement: (1) personality-based disagreement, beget by ideology, venality, or incompetence of the experts themselves, (2) judgement based disagreement, where information gaps exist, or (3) structural disagreement generated by different problem definitions or organizing principles held by experts [13].

In the clinical domain, expert disagreement is prevalent in diagnostic tasks that rely on visual analysis of subjective criteria. EEG interpretation is one such example. For instance, in the context of epilepsy diagnosis, interictal epileptiform discharges (IEDs) are key distinguishing features within an EEG—however, neurologists will often disagree over whether a particular waveform constitutes an IED. In a study by Bagheri et al. (2017), it was found that inter-rater agreement rates surrounding IED detection could be predicted based on particular wavelet features, given that the sample of experts was large enough [1]. In the context of sleep stage classification, which also depends on the identification of transient and infrequent features (i.e., sleep spindles and K-complexes), the average agreement rate among experts is 82.6% [19].

2.2 Adjudication in Medical Data Analysis

Disambiguating edge cases generated by inter-rater disagreement in the medical domain is a matter of contention in the contemporary literature. Majority-vote techniques have been criticized for their tendency to promote artificial consensus over valuable data or insights that might be had from group discussion and deliberation [23]. Indeed, group deliberation—where group members who hold conflicting beliefs present arguments and weigh evidence in light of their individual positions in order to reach a decision—has been shown to be a useful and productive technique for aggregating expert opinions and reaching consensus.

A study by Krause et al. (2017) found that in-person, group deliberation resulted in significantly higher recall among experts in diagnosing eye disease from images of the fundus, when compared to the majority vote technique [10]. Furthermore, it was demonstrated in the same study that group deliberation, when performed on just a small portion of a dataset, can be used to train the hyperparameters of deep learning models for more effective automated analysis. Guan et al. (2018) later used the same consensus data set to train multiple, grader-specific machine learning models, and showed that the aggregate performance of these models could beat out a single-prediction model trained with majority labels [8].

Adjudicated diagnoses have also proved valuable as reference standards for training machine learning models. In work done by Rajpurkar et al. (2017), cardiologists engaged in group deliberation to generate an adjudicated electrocardiogram (ECG) data set in the context of arrhythmia detection. This consensus validation data set was then used as a benchmark for a convolutional neural network, which was found to outperform individual cardiologists in ECG classification when trained solely on independent data labels [16].

Where sleep staging is concerned, it has been argued that group deliberation, also referred to as "consensus-scoring" or adjudication, is an optimal method of training human sleep scorers [15].

2.3 Computational Models of Argumentation

Argumentation is an approach to reasoning focused not only on the conclusions reached, but also on the data and the inference steps involved in inferring conclusions from the data. Argumentation has a considerable history in the field of computer science, including the problem of understanding common patterns of argumentative discourse in human decision making (e.g., [6, 25]) mapping natural language to a more formal, machine readable representation of argumentative discourse (e.g., [3–6, 11, 12, 18]), and using formal representations of arguments to generate new conclusions for previously unseen queries (e.g., [14, 17]).

3 APPLICATION DOMAIN

We leverage biomedical time series classification, a field with typically low inter-scorer reliability, as an application domain for embedding our work. In particular, we use examples from sleep stage classification, the expert task of mapping a sequence of fixed-length pages (typically 30 seconds) of continuous multimodal medical time

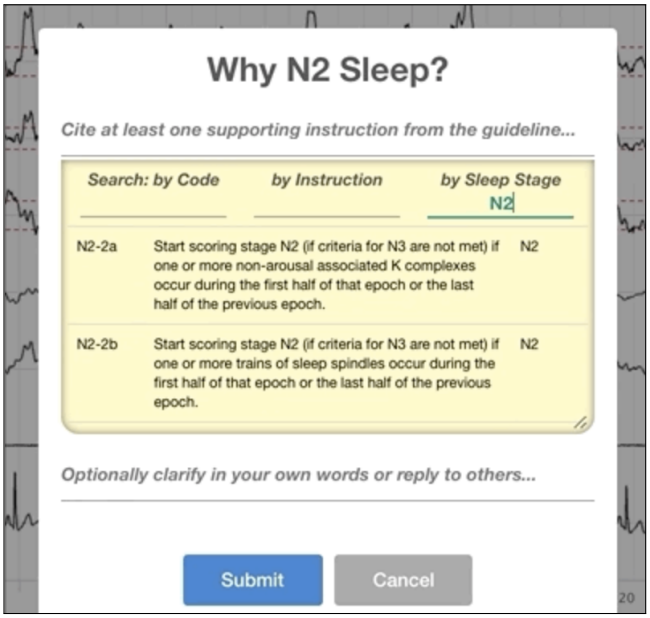


Figure 2: Rationale form for expert graders to cite guideline instructions in support of their classification decision.

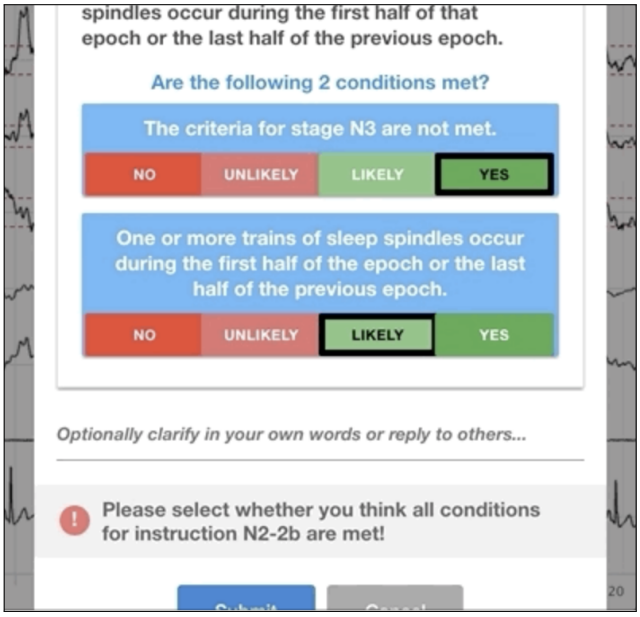


Figure 3: Expert graders specify their level of confidence for individual conditions required for a cited instruction.

series (*polysomnogram*, see Figure 1) to a sequence of discrete sleep stages (*hypnogram*). Each fixed-length page of time series (epoch) is classified into one of five different stages of sleep—Wake, NREM1, NREM2, NREM3 or REM sleep—based on the stage comprising the greatest portion of the epoch. Rosenberg and van Hout [19] conducted a study on inter-scorer reliability in sleep stage classification, finding that expert agreement averages around 82.6%.

4 PROPOSED SOLUTION

We propose to augment the traditional process of collecting ground truth labels for supervised learning (i.e., querying one or more experts for the "correct" label to a given input example), by introducing an extra step to elicit the reasons for certain classification decisions (i.e., *rationale*) in a structured form. In particular, we propose to collect expert rationale in the form of propositional logic where experts specify the evidence and inference rules they used to arrive at their classification decisions. In this section, we guide the reader through the input and output of our proposed approach and the intermediate steps required to transform the input to the output. We illustrate our explanations with examples from sleep stage classification.

Input. Our system will take as input a pool of *human experts* and a set of *data objects* (e.g., images, text documents, medical time series) to be classified into one of several preset *categories*. In the case of sleep stage classification, a set of pages of physiological time series (see Figure 1) is classified into one of 5 stages of sleep by a pool of sleep technologists.

Output. For each input data object, our method will output a distribution of classification labels, one from each expert. For those data objects that led to some disagreement among the experts throughout the labeling process, the system will also output each individual expert's rationale for their final classification decision in the form of propositional logic. In other words, for ambiguous cases, the system will list the inference rule(s) each expert used to arrive at a certain classification decision as well as the expert's confidence levels for the evidence criteria that need to be met in order for the chosen inference rule(s) to be applicable. An example for sleep stage classification may look as follows:

- **Expert A:**
  - **Classification:** Wake
  - **Rule:** "*[W-3a] Score epochs without alpha rhythm as stage W if eye blinks are present.*"
  - **Confidence for Evidence Criteria:**
    - Alpha rhythm absent: **Yes**
    - Eye blinks present: **Likely**
- **Expert B:**
  - **Classification:** NREM1 Sleep
  - **Rule:** "*[N1-2] In patients who generate alpha rhythm, score stage N1 if the alpha rhythm is extenuated and replaced by low-amplitude, mixed-frequency activity for more than half of the epoch.*"
  - **Confidence for Evidence Criteria:**
    - Patient generates alpha rhythm: **Likely**
    - Alpha rhythm is extenuated and replaced by low-amplitude, mixed-frequency activity for more than half of the epoch: **Yes**

4.1 Capturing Structured Rationale

Our solution requires two components to capture rationale in the format above:

(1) Rule-based representation of the classification guidelines
(2) User interface to collect rationale in structured form

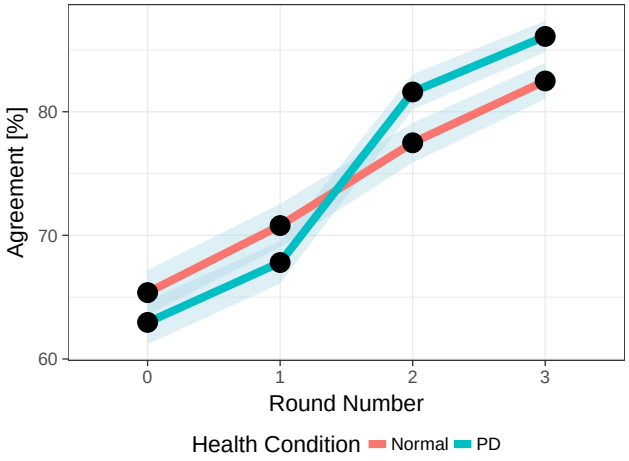


Figure 4: Agreement rate by adjudication round number and patient's health condition. Transparent overlays indicate 95 percent confidence intervals.

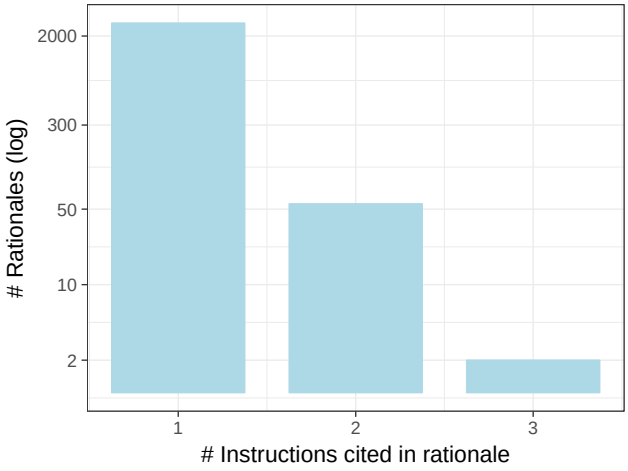


Figure 5: Number of rationales citing one, two and three guideline instructions.

Rule-based representation of the classification guidelines. The first step in collecting rationale in a structured form is to define the set of possible inference rules for classifying objects, and a set of evidence criteria (i.e., Boolean propositions) that need to be met in order for a given rule to be applicable. For our application domain of sleep stage classification, we adapted the official *AASM Manual for the Scoring of Sleep and Associated Events* [9] into a set of 36 individual inference rules (8 for Wake, 10 for NREM1 Sleep, 9 for NREM2 Sleep, 3 for NREM3 Sleep, and 6 for REM Sleep) with 52 unique evidence criteria.

User interface to collect rationale in structured form. The resulting rule-based representation of the classification guidelines needs to be exposed through a user interface enabling experts to specify individual inference rules and to indicate the extent to which they believe that the evidence criteria required for the selected rules are met. Figure 2 illustrates our implementation of such an interface. In our example, the interface starts out as an empty rationale form with two input fields—one to select discrete inference rules, and one to optionally explain more in one's own words. The first input field will automatically suggest possible inference rules based on the current classification decision (e.g., Wake) and the keywords typed into the input field. Once the user selects an inference rule, the rationale form automatically lists the evidence criteria that need to be true in order for the rule to be applicable (Figure 3), prompting the user to indicate the extent to which they believe each condition is met, one of: No, Unlikely, Likely, Yes. The interface produces warnings for invalid inputs (e.g., selecting rules while indicating that their conditions are not met) to ensure the user has selected at least one inference rule in support of their classification decision and specified their confidence level for each of the evidence criteria, before submitting the rationale and proceeding to the next disagreement case.

5 PILOT EXPERIMENT

A pilot experiment was conducted to explore the usefulness of the proposed procedure and to demonstrate sample analyses made possible by the resulting structured adjudication data. For our pilot experiment, we sampled six EEG recordings from six unique patients, three with Parkinson's disease (PD) and three normal control subjects. As is the case for other neurological disorders, sleep studies from PD patients may exhibit slight differences in the expression of sleep-relevant features (e.g., sleep spindles and K-complexes), compared to healthy subjects, and may therefore lead to different disagreement patterns among domain experts. For annotation, we recruited 18 sleep technologists as expert graders forming six panels of three experts each. Each EEG recording was assigned to exactly one of the six expert panels, and each expert grader participated in exactly one panel. Graders first performed an initial independent round of scoring on their assigned recordings, followed by three rounds of adjudication, one round per grader in the panel. In each adjudication round, the active grader stepped through each individual epoch with any level of disagreement among panel members, re-scored the epoch and provided a rationale for their reclassification decision. In each adjudication round and for each disagreement epoch, the active grader was presented with the most recent grades from all three panel members, as well as the grades and rationales submitted during each of the preceding rounds. The three panels adjudicating recordings from PD patients used the structured, guideline-centric way of collecting rationale during adjudication discussions. In contrast, the three remaining panels adjudicating recordings from normal control subjects used the non-structured, free-form way of collecting rationale.

6 RESULTS

The output of our rationale form (i.e., sleep stage classification labels and their given expert rationales) provided the basis for a detailed set of quantitative results across multiple items of analysis.

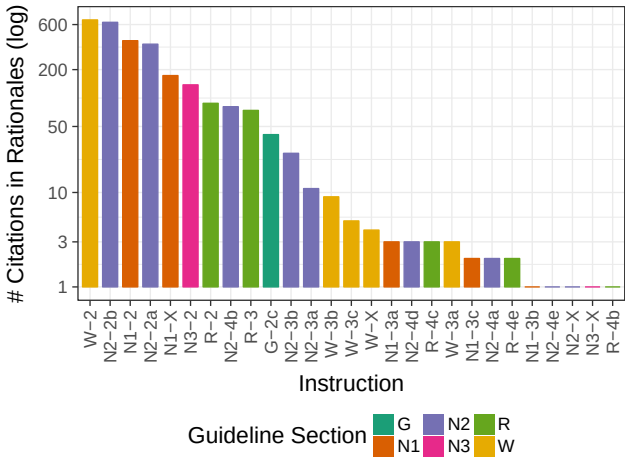


Figure 6: Number of citations per guideline instruction. Bar colours indicate the guideline section for each instruction: General (G), NREM1 Sleep (N1), NREM2 Sleep (N2), NREM3 Sleep (N3), REM Sleep (R) and Wake (W).

For ambiguous cases in which experts were required to provide a rationale and confidence levels for their assessments, data were collected on the number of citations each guideline instruction in the AASM scoring manual [9] received. As seen in Figure 6, select guideline sections received a disproportional number of citations compared to others, based on the sleep stage to which the classification label pertained. The guideline section listing rules for scoring NREM2 sleep received the greatest number of citations overall, while instruction W-2 received the highest number of citations for a given instruction. Certain guideline instructions, such as R-4b, received only a single citation across all ambiguous cases.

Despite a wide distribution in the subsections indexed by expert rationales, rarely did individual rationales cite more than one guideline instruction (see Figure 5). The significant majority of rationales (>2000) referenced only one instruction, while less than 3% of rationales given cited two. For all rationales collected, none cited more than three instructions in total.

Since adjudication decisions in sleep stage classification often hinge on the presence or absence of distinguishing features of an EEG waveform, references to feature types were also collected from expert rationale output. 15 basic features in total were mentioned across all rationales, again, with select features (i.e., alpha rhythm, train of sleep spindles, non-arousal associated K-complexes) receiving far more mentions than others. However, unlike the data for the number of citations per guideline instruction (Figure 6), there was a much flatter distribution in the number of mentions received by each feature type, with no substantial difference between the number of mentions across 7 of the 15 feature types, as shown in Figure 7.

Results from three adjudicated rounds of scoring with 6 panels of expert participants (3 experts per panel) showed a marked decrease in the number of controversial cases—and thus an increase in inter-rater agreement—between independent annotation and the end of adjudication, as seen in Figure 4. For the 3 panels that adjudicated EEG recordings from healthy controls, average agreement rate increased across three rounds of adjudication, and rose from ~66% to ~83%. These 3 panels provided rationales for disagreement cases through a non-structured, free-form interface. Among the three panels that adjudicated EEG recordings from Parkinsonian patients, average inter-rater agreement rate increased across three rounds of adjudication from ~63% to ~86%. Here, experts provided classification rationales for disagreement cases through our structured, guideline-centric interface. While overall change in average agreement rate did not substantially differ between free-form and structured adjudication, there were clear differences between the rates at which agreement rates increased across adjudication rounds in these two scenarios. Where free-form rationales were provided for disagreement cases between experts, average inter-rater agreement increased in a linear fashion between independent annotation and adjudication. For those expert panels that provided their rationales through the structured, guideline-centric interface, inter-rater agreement increased in a step-wise fashion, with the biggest jump in average agreement rate occurring between the first and second round of adjudication.

Furthermore, a stepwise logistic regression model was used to understand which feature types, when mentioned during a given adjudication round, were associated with the probability that the active adjudicator will change their classification decision in the same round. As outlined in Table 1, the likelihood that an expert would change their classification decision could be predicted based on which feature types were mentioned in their rationale. Graders mentioning arousals ($p < 0.05$) or low-amplitude mixed frequency activity (LAMF; $p < 0.001$) in their adjudication rationales were significantly more likely to stick with their classification decision than those not mentioning these features. Citing instructions pertaining to K-complexes ($p < 0.01$) or trains of sleep spindles ($p < 0.001$) was significantly associated with a change in the grader's classification decision in the same round. The logistic model selected additional feature types—low chin EMG tone, reading eye movements and rapid eye movements (REM)—contributing to the model fit without statistical significance.

7 DISCUSSION

The main contribution of our present work is a structured system and procedure for capturing expert rationale during adjudication of complex data sets–in this case, within the application domain of sleep stage classification. In addition to querying a group of experts for data labels (i.e., a sleep stage for a given epoch), we solicited rationales from these expert graders in the form of propositional logic (i.e., the reason for their classification decision in the form of a sleep scoring guideline), as well as their self-reported confidence levels for their evidence criteria. During structured adjudication, average inter-rater agreement rose by roughly 23%.

Applications of this system exist both within and outside of the present application domain. In the context of sleep stage classification, controversial cases surrounding classification decisions—especially those that remain after adjudication—may point to ambiguous instructions in the sleep scoring guidelines. With output from a structured rationale form, specific guideline sections and

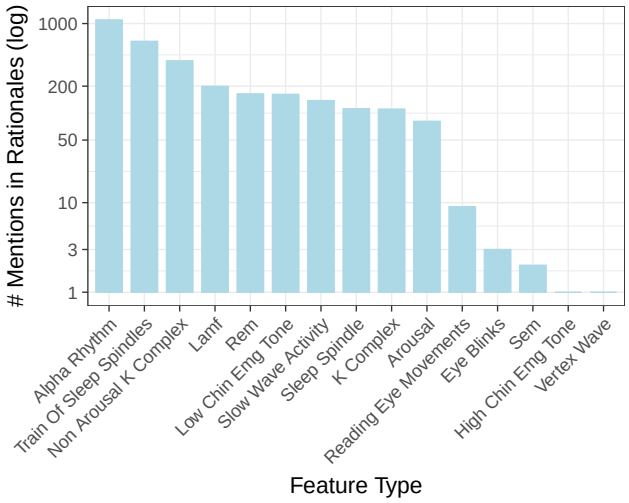


Figure 7: Number of times each feature type was mentioned in a rationale.

Table 1: Logistic model for understanding the likelihood of a grader changing their decision in a given round, based on the feature types mentioned in their rationale.

	Model Parameters			
Variable	$\hat{\beta}$	*Std. Error*	*t*	*p-value*
K Complex	0.93	0.33	2.77	**
Train Of Sleep Spindles	0.37	0.09	3.97	***
Arousal	-0.52	0.25	-2.10	*
Lamf	-1.72	0.51	-3.36	***
Low Chin Emg Tone	16.28	624.19	0.03	
Reading Eye Movements	-14.22	294.25	-0.05	
Rem	-14.22	624.19	-0.02	

instructions can be indexed for potential iteration. Likewise, particular feature types mentioned in given rationales may require more rigorous definitions.

Beyond the present application domain, we have shown that the use of statistical models like the one illustrated in Table 1 can help predict outcomes of structured adjudication, e.g., the likelihood that an individual grader will change their assessment based on particular evidence criteria. Similar models could be used to predict how many rounds of adjudication are necessary to resolve ambiguous cases. Of the total ~23% rise in average inter-rater agreement during our structured adjudication procedure, ~19% of that increase occurred between the first and second adjudication rounds. Additional analysis may reveal patterns as to which types of disagreements are resolved early vs. late.

The most notable limitation of work of this kind lies with the fact that adjudication is costly–especially in terms of time investment from expert graders. In addition, adjudication procedures like the one we deployed in this application domain depend on the existence of standardized grading guidelines (like the AASM sleep scoring manual [9]) which must be agreed upon by all expert graders participating. In this study, we mapped an existing standard scoring manual into a set of scoring instructions integrated into the adjudication interface. In cases where a single agreed-upon grading guideline does not exist in the community (e.g., interpretation of EEGs for epileptiform abnormalities), there is interesting potential for future work in the iterative development of scoring guidelines based on adjudication procedures that become increasingly structured, and thus less time-intensive, over time.

This study has laid the groundwork for future work in the area of analyzing the sources of inter-rater disagreement in labelling complex datasets, within the application domain of sleep stage classification. Throughout the continuation of this project, we intend to prepare and make publicly available a high-quality dataset of adjudicated human polysomnograms. The use cases for this data are severalfold. First, there is the potential to derive concrete suggestions for guideline development. Second, the data could be used to build machine learning models for predicting ambiguity and sources of disagreement for previously unseen data, an ability that would be helpful for guiding human labelling resources, and for establishing new pathways towards more informed and explainable concepts of uncertainty in machine learning. Third, these adjudication data sets can be used to train human graders in better disambiguating edge cases by leveraging structured information from adjudication rounds to target grader training towards different categories of ambiguity.

8 CONCLUSION

In this work, we introduced a novel perspective on the problem of handling expert disagreement in ambiguous classification tasks by proposing a structured procedure for collecting expert arguments put forward during panel-based adjudication in the form of propositional logic. We demonstrated the applicability of our approach in the context of medical time series analysis for sleep stage classification, and showcased how the data produced can facilitate detailed quantitative analyses of discussion contents and outcomes. Our solution has implications for the broader field of supervised learning from human-labeled data by translating the problem of trustworthy AI systems to that of trusted and explainable ground truth.

ACKNOWLEDGMENTS

We thank Rui DeSousa for his invaluable help in recruiting participants for this study. This work was funded by NSERC CHRP (CHRP 478468-15) and CIHR CHRP (CPG-140200).

REFERENCES

[1] Elham Bagheri, Justin Dauwels, Brian C. Dean, Chad G. Waters, M. Brandon Westover, and Jonathan J. Halford. 2017. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology* 128, 10 (10 2017), 1994–2005. https://doi.org/10.1016/j.clinph.2017.06.252

[2] John Beatty and Alfred Moore. 2010. Should We Aim for Consensus? *Episteme* 7, 3 (2010), 198âĂŞ214. https://doi.org/10.3366/E1742360010000948

[3] Floris Bex, Henry Prakken, Chris Reed, and Douglas Walton. 2003. Towards a Formal Account of Reasoning about Evidence: Argumentation Schemes and Generalisations. *Artificial Intelligence and Law* 11, 2/3 (2003), 125–165. https://doi.org/10.1023/B:ARTI.0000046007.11806.9a

[4] Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards Argument Mining from Dialogue. In *Computational Models of Argument - Proceedings of {COMMA}*

2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014. 185–196. https://doi.org/10.3233/978-1-61499-436-7-185

[5] Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The Knowledge Engineering Review* 21, 04 (12 2006), 293. https://doi.org/10.1017/S0269888906001044

[6] Robin Cohen. 1987. Analyzing the Structure of Argumentative Discourse. *Comput. Linguist.* 13, 1-2 (1 1987), 11–24. http://dl.acm.org/citation.cfm?id=26386.26388

[7] Luciana Garbayo. 2014. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making* 539 (2014), 35–45. http://link.springer.com/10.1007/978-3-319-04280-0%5C_5

[8] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*. https://arxiv.org/pdf/1703.08774.pdf

[9] Conrad Iber, Sonia Ancoli-Israel, Andrew L Cheeson Jr., and Stuart F Quan. 2007. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine.

[10] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* (3 2018). https://doi.org/10.1016/j.ophtha.2018.01.034

[11] John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining at ACL 2015*. 127–136. https://doi.org/10.3115/v1/W15-0516

[12] John Lawrence and Chris Reed. 2016. Argument Mining using Argumentation Scheme Structures. *Proceedings of the 6th International Conference on Computational Models of Argument (COMMA 2016)* 0 (2016), 379 – 390. https://doi.org/10.3233/978-1-61499-686-6-379

[13] Jeryl L. Mumpower and Thomas R. Stewart. 1996. Expert Judgement and Expert Disagreement. *Thinking & Reasoning* 2, 2-3 (7 1996), 191–212. https://doi.org/10.1080/135467896394500

[14] Simon Parsons, Elizabeth Sklar, Jordan Salvit, Holly Wall, and Zimi Li. 2013. ArgTrust: Decision Making with Information from Sources of Varying Trustworthiness. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '13)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1395–1396. http://dl.acm.org/citation.cfm?id=2484920.2485242

[15] Thomas Penzel, Xiaozhe Zhang, and Ingo Fietze. 2013. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *Journal of Clinical Sleep Medicine* 9, 1 (2013), 81–87.

[16] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. (7 2017). http://arxiv.org/abs/1707.01836

[17] Chris Reed and Timothy Norman. 2004. *Argumentation Machines*. Argumentation Library, Vol. 9. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-017-0431-1

[18] Chris Reed and Doug Walton. 2005. Towards a Formal and Implemented Model of Argumentation Schemes in Agent Communication. 19–30. https://doi.org/10.1007/978-3-540-32261-0{_}2

[19] Richard S. Rosenberg and Steven van Hout. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* (1 2013). https://doi.org/10.5664/jcsm.2350

[20] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'18)*. New York City, NY. https://doi.org/10.1145/3274423

[21] Mike Schaekermann, Edith Law, Kate Larson, and Andrew Lim. 2018. Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing at HCOMP 2018*. Zurich, Switzerland.

[22] Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. 2016. Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*. San Jose, CA.

[23] Miriam Solomon. 2006. Groupthink versus The Wisdom of Crowds : The Social Epistemology of Deliberation and Dissent. *The Southern Journal of Philosophy* 44, S1 (3 2006), 28–42. https://doi.org/10.1111/j.2041-6962.2006.tb00028.x

[24] Miriam Solomon. 2007. The social epistemology of NIH consensus conferences. In *Establishing medical reality*. Springer, 167–177.

[25] D Walton, C Reed, and F Macagno. 2008. *Argumentation Schemes*. Cambridge University Press. https://books.google.ca/books?id=qc3LCgAAQBAJ

Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing

Rafael Zequeira Jiménez
Technische Universität Berlin
Berlin, Germany
rafael.zequeira@tu-berlin.de

Anna Llagostera
Rohde & Schwarz SwissQual AG
Zuchwil, Switzerland
anna.llagostera@rohde-schwarz.com

Babak Naderi
Technische Universität Berlin
Berlin, Germany
babak.naderi@tu-berlin.de

Sebastian Möller
Technische Universität Berlin
DFKI Projektbüro Berlin
Berlin, Germany
sebastian.moeller@tu-berlin.de

Jens Berger
Rohde & Schwarz SwissQual AG
Zuchwil, Switzerland
jens.berger@rohde-schwarz.com

ABSTRACT

Crowdsourcing is a great tool for conducting subjective user studies with large amounts of users. Collecting reliable annotations about the quality of speech stimuli is challenging. The task itself is of high subjectivity and users in crowdsourcing work without supervision. This work investigates the intra- and inter-listener agreement withing a subjective speech quality assessment task. To this end, a study has been conducted in the laboratory and in crowdsourcing in which listeners were requested to rate speech stimuli with respect to their overall quality. Ratings were collected on a 5-point scale in accordance with the ITU-T Rec. P.800 and P.808, respectively. The speech samples were taken from the database ITU-T Rec. P.501 Annex D, and were presented four times to the listeners. Finally, the crowdsourcing results were contrasted to the ratings collected in the laboratory. Strong and significant Spearman's correlation was achieved when contrasting the ratings collected in both environments. Our analysis show that while the inter-rater agreement increased the more the listeners conducted the assessment task, the intra-rater reliability remained constant. Our study setup helped to overcome the subjectivity of the task and we found that disagreement can represent a source of information to some extent.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; *Speech / audio search*; • **Human-centered computing** → **User studies**; **Laboratory experiments**; **Computer supported cooperative work**; *Empirical studies in HCI*; • **Computing methodologies** → Speech recognition.

KEYWORDS

inter-rater reliability, speech quality assessment, crowdsourcing, listeners' agreement, subjectivity in crowdsourcing

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317084

ACM Reference Format:
Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller, and Jens Berger. 2019. Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317084

1 INTRODUCTION

In the recent years, crowdsourcing has become a convenient approach for solving a multitude of problems that require human input. Crowdsourcing (CS) can be understood as the parallelization of computational work in programming environments, which normally consist of segmenting the work into multiple small and independent tasks. These small tasks, are accomplished by a diverse pool of users in exchange for a monetary compensation. This approach has been adopted in multiple domains and is specially beneficial for conducting subjective user studies.

The quality of the transmitted speech signal is of main importance for telecommunication network providers, as it is one of the main indicators to evaluate their systems and services. The speech signal can be damage by the codecs, linear and non-linear filters, bandwidth limitations, and other elements as reported in [20]. Technological advances within traditional and modern packed-based (Voice-over-IP) telephony networks, introduces new codecs. This demands new empirical subjective user studies to understand how end users perceive these impairments in the speech signal.

Traditionally, subjective speech quality studies has been carried out in Laboratory (Lab) environments under controlled conditions and with professional audio equipment. This way a good control over the experiment setup and a proper control over the participants can be achieved but with some mayor disadvantages, e.g. it is expensive, time consuming, and often the number of participants is rather low. Therefore, sometimes the results might not be representative of a larger population. In contrast, CS permits to reach a fairly demographic distributed pool of users at a fraction of the cost and time.

The study of intra- and inter-rater reliability in speech quality experiments in CS is rather poor. Often, workers evaluate just a portion of the dataset which makes it difficult or even impossible to compute the intra- and inter-rater agreement. This decision

is normally made so the test session can be kept short [7] while avoiding the workers' boredom [21]. This was the case in [18] and in [19], where authors carried a speech quality assessment experiment in a crowdsourcing platform, and workers could assess just the 2.5% of the dataset every time they participated in the study.

Also, work in [25] used CS to investigate the quality of speech stimuli. In the presented experiment, only two workers (out of more than 200) evaluated the entire database and thus, no analysis was given regarding the intra- or inter-rater reliability. Moreover, research in [22] used Amazon Mechanical Turk to investigate the naturalness of synthesized speech in a discrete 5-point scale. Again, workers were able to evaluate just the 3.1% of the available files each time they participated in the study. However, authors repeated the experiment but only to prove the validity of the framework they proposed.

In this work, listeners had the chance to evaluate four times the speech samples in the dataset, so we collected enough data to analyze the agreement among and within subjects. Our hypothesis is that the between and withing listener agreement would increase from the first to the last time the users execute the listening test, as well as the accuracy of the quality scores.

Moreover, authors of [8] proposed the "SOS hypothesis" as an alternative for auditing the reliability of the test results, and to measure consistency in subjective Quality of Experience (QoE) studies. This hypothesis, models a square relationship between the standard deviation of the opinion scores (SOS) and the Mean Opinion Scores (MOS). The authors proposed it also for measuring comparability across multiple QoE studies in CS.

To determine whether our hypothesis is true, we focused on the subjective evaluation of new state-of-the-art codecs (EVS and Opus) in different types of transmission, e.g. VoLTE, circuit switched mobile and VoIP applications like WhatsApp. This evaluation was both carried in the laboratory and in crowdsourcing following international standards, e.g. ITU-T Rec. P.800 [10] and P.808 [12], respectively. Participants in both environments listened four times all of the speech stimuli in the dataset, and gave their opinion about the overall quality in a 5-point scale. The crowdsourcing results are then contrasted to the Lab and we analyze the intra- and inter-listener agreement. Additionally, we investigate the correlation to the Lab results per crowd-worker at the different test session, in order to determine at which one the quality scores are more accurate, which would help to save resources in future studies.

The remainder of this paper is structured as follows. The study setup in the laboratory as well as in the CS environment is outlined in Section 2. Section 3 presents the main findings, firstly the CS results are contrasted with the Lab, and then the agreement between and within subject is analyzed. Section 4 summarizes our findings and concludes.

2 METHOD

In the following, the speech material employed in our investigation is presented as well as the study conducted in the Lab. This latter, aimed at evaluating the user perception of todays' mobile telephony services as VoLTE, circuit switched mobile (GSM and UMTS) and VoIP OTT applications like WhatsApp. Finally, the CS experiment is detailed which intended to replicate the results gathered in the Lab.

2.1 Database

The database consists of a set of real-field and offline samples at different audio bandwidths from below narrowband and up to fullband. The distribution among the speech stimuli is as follows: 21% narrowband (NB), 50% wideband (WB), 23% super-wideband (SWB) and 6% full-wideband (FWB).

The real-field recordings (71% of the speech samples) in the database were collected in Switzerland during September 2018, under good, average and bad coverage network conditions, and using Rohde & Schwarz SwissQual equipment. It includes state-of-the-art measurements such as:

- VoLTE calls with EVS at 24.4 kbit/s SWB
- WhatsApp calls in LTE with Opus at 20 kbit/s WB
- 3G mobile to mobile calls with AMR-WB at 12.65kbit/s and at 23.85 kbit/s
- 3G mobile to mobile calls with AMR-NB at 12.2 kbit/s
- 3G/2G mobile to mobile calls with transcoding from AMR-WB at 12.65 kbit/s to AMR-NB at 12.2 kbit/s

The offline (29% of the speech samples) coded conditions in the dataset aims at replicating those common conditions that can be seen in the field. The three fullband conditions in the test were a fullband reference and two anchors with packet loss. Four additional low quality simulated conditions were obtained by either adding packet loss to some codec conditions, or re-encoding the reference sample multiple times with the same settings. The sample used was the composed female/male German sample from the ITU-T Rec. P.501 Annex D [9]. In total, 53 speech stimuli, 7 seconds long on average, are arranged accounting for 53 degradation conditions.

2.2 Laboratory Study

The study was carried at SwissQual's listening Lab in October 2018. 24 native German listeners (11 female and 13 male) were invited individually to conduct the test. As previously pointed out, this P.800 [10] ACR listening test, targeted the subjective evaluation of speech stimuli encoded with state-of-the-art codecs (e.g. EVS [3], AMR-WB, AMR-NB [1, 2] and Opus [1]) under ideal and live good/average/bad coverage conditions.

Each listener evaluated the quality of the 53 speech stimuli in a five-point absolute category rating (ACR) quality scale with the options (translated from German): "Excellent", "Good", "Fair", "Poor" and "Bad". The presentation order was randomized, and to achieve small enough confidence intervals, each listener evaluated four times the 53 speech samples.

Prior the assessment, participants went through a training in which they judged five stimuli and they become familiar with the test setup. The speech samples were presented diotically to the subjects by means of diffuse field equalized headphones (Grado SR 60). The presentation level was 73 dB(A) SPL at each ear (equivalent to -26dB OVL). More information about the Lab study can be found in [4].

[1] https://opus-codec.org/ last accessed March 2019

All in all, subjective quality assessments to the 53 stimuli were gathered, made by 24 different listeners. The Mean Opinion Score (MOS) [11] was computed for each of the stimulus by averaging the subjective ratings given by all of the subjects to the same speech degradation condition. These MOS scores are then taken as a reference for the analysis presented in this work, from now on referred as “Lab-MOS".

Additionally, a Kendall’s coefficient of concordance (W) [13] was run to determine if there was agreement among the listeners’ ratings. This test revealed a statistical significant agreement across the participants when they assessed all of the speech stimuli, $W = 0.86, p < 0.001$. The reliability of the data is verified when high agreement exists in the ratings that different participants provide to the same speech stimulus. This high Kendall’s coefficient, exposes a low variability across the individual ratings, which demonstrate a high confidence of the collected mean opinion scores.

2.3 Crowdsourcing Study

The goal of the Crowdsourcing (CS) study was to check whether the Lab result could be reproduced in a CS environment. Additionally, we wanted to verify whether the listeners in CS would achieve the same level of agreement considering the subjectivity of the task. With that purpose, we keep constant some important settings from the Lab, e.g. targeting the study to native Germans, gathering at least 96 ratings per sample and conducting the experiment complying with the ITU-T Rec. P.800 [10].

As a CS platform, we used clickworker [2], which is based in Germany as well as most of its users, therefore a good fit for our experimental needs.

The study setup was similar to that one of [25], and with the guidelines of the ITU-T Rec. P.808 [12] in mind, which has been proven to produce good results [17]. Firstly, a screening task was used to collect basic demographic information and to check the workers’ German command, so we could invite to the study only those reporting a native level. For this, a few short German audio-passages were arranged, and the workers were requested to select the right affirmation out of multiple options that were available.

Secondly, the invited workers went through a training phase before they could participate in the speech quality assessment task (SQAT) and evaluate the database. This training permitted to control the headphones’ two-eared usage. A short math exercise was prepared for this with digits panning left to right in stereo [18]. This question allowed to ban the sloppy workers from participating in the SQAT during one hour. Additionally, workers were presented with five stimuli for anchoring, so they could get to know what to expect during the assessment while becoming familiar with the interface. The stimuli employed here were the same as the ones used for anchoring in the Lab study.

Finally, when workers finished the training properly, they were presented automatically with the SQAT. Then, listeners were requested to rate the overall quality of 53 speech stimuli in a five-point scale, see Figure 1. Workers could execute the SQAT up to four times (like in the Lab). Also, one hour timeout was set after which they

[2]https://www.clickworker.com last accessed March 2019

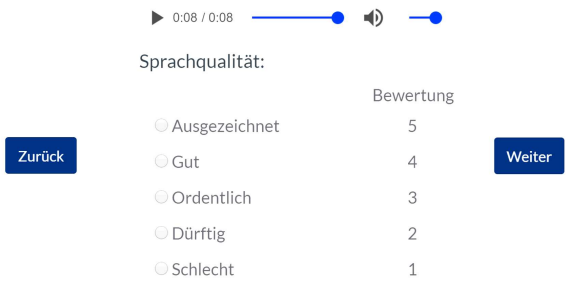


Figure 1: Graphical interface presented to the workers for the SQAT (in line with [12]). The text translate from German: “Speech Quality" and “Rating". The scale (in descending order): “Excellent", “Good", “Fair", “Poor" and “Bad".

were forced to conduct once more the training [19]. To ensure quality, one trapping question (TQ) was inserted randomly within the first five stimuli of every ten speech samples. The TQs’ GUI was the same like in the rest of stimuli, but the audio was modified to highlight the importance of the listeners’ work, and to request them to select an specific option on the rating scale, so they could prove they were conducting the task conscientiously [18]. As result, workers were presented with 58 speech samples in total.

Listeners were prevented from conducting the assessment during one hour, when they missed one of the trapping questions. In this case, the ratings from the set of those ten stimuli were labeled as unreliable. Additionally, at the end of the SQAT workers were requested to express in a slider how tired they felt, 1 was (translated from German): “not exhausted at all" and 11 “extremely exhausted".

To do a “direct" comparison to the Lab results, we wanted to accomplish that at least 24 crowd-workers would execute the SQAT four times, therefore we used a bonus system to motive the participation. Listeners received 1.00 EUR for each time they completed the SQAT, and 0.40 EUR (extra payment) the second, third and fourth time they conducted the SQAT.

3 RESULTS

52 workers in total (100% native Germans, 51.9% female, 96.2% from Germany, balanced age) produced 8321 ratings. Only three workers failed some trapping questions in the SQAT. Two of them missed all of the TQ, while the other worker missed just the last one. Then, a total of 119 ratings were discarded.

To assess whether the ratings gathered in CS correlate to the ones collected in the Lab, we calculated the Spearman’s rank-order correlation. This metric provides a measure of the strength and direction of the association and/or relationship between two continuous or ordinal variables. We then determined the Spearman’s correlation between the Laboratory ratings (Lab-MOS), and the ratings in CS given by the workers that conducted the SQAT four times like in the Lab. 29 workers in total participated four times and yielded 6148 assessments. These subjective ratings were averaged by degradation condition in order to compute the MOS scores, which we refer to as “CS-MOS". Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of

a scatter-plot. Additionally, the Root Mean Square Error (RMSE) between the Lab-MOS and the CS-MOS was calculated. There was a statistically significant, strong positive correlation between the Lab-MOS and the CS-MOS, $\rho = 0.978$ ($p < .001$), as well as a low $RMSE = 0.441$. This result, indicates the validity of CS as a tool for collecting reliable annotations about the quality of speech stimuli, regardless of how close the different degradation conditions may be to each other.

3.1 Inter-rater Reliability

To assess further the validity of the collected ratings, we investigated the level of agreement among the listeners in each of the four stages in which they conducted the SQAT. We wanted to verify whether the agreement fluctuated from the first to the fourth repetition and also if it varied with respect to the Spearman's correlation and the RMSE.

The Kendall's coefficient of concordance (W) [13] can be used as such a measure of inter-rater agreement for continuous and ordinal variables when there are two or more raters [5, 15]. For this analysis we considered the single rating given by the 29 workers to the 53 stimuli at each of the repetitions. Hence, four Kendall's W coefficients were calculated on this ordinal data.

Furthermore, we computed the correlation and RMSE between the Lab-MOS and the ratings provided by the 29 crowd-workers averaged per file, at repetition one, two three and four. Table 1 outline these results. While the RMSE and the correlation remained almost constant, the agreement between workers increased slightly from repetition one to four (e.g. 0.67 to 0.68). This is understandable since a listeners would become more confident with the ratings he or she provides, the more they participate in the study. However, it can be seen that the lowest agreement ($W = 0.65$) was achieved during the third repetition. To investigate further into this, we then conducted a two-way mixed ANOVA [16, 24]. The analysis showed that indeed the main effect of repetition presented a statistically significant difference in the mean ratings at the third repetition, for three of the speech degradation conditions in the dataset, e.g. $C07$, $C15$ and $C37$. Results are presented in Table 2.

We can conclude that the decrease in the between listener agreement at the third repetition, facilitated to identify those three conditions that apparently were the most difficult for the listeners to assess. This outcomes suggest that to gather reliable speech quality annotations for these degradations conditions, the assessment task should be addressed to a large pool of listeners, so the confidence intervals of the quality ratings could be lowered down. In any case, further investigation would be needed to determine the reasons for the differences in the ratings of these speech conditions.

All in all, the high Kendall's coefficient achieved indicate that the SQAT was well designed with a low ambiguity. And we can assume that, all of the workers understood the instructions, and thus, most of the variance across the ratings can be explained by the differences between the speech degradation conditions, and not by individual differences in the listeners' evaluations (e.g. due to different test interpretations). This outcome confirm the reliability of the collected ratings.

Table 1: Kendall's (W) coefficient of agreement, Spearman's correlation (ρ) and RMSE between the Lab-MOS and the CS-MOS for each of the times the workers conducted the SQAT.

Repetition	W	ρ	$RMSE$
1	0.6712^*	0.9774^*	0.4452
2	0.6673^*	0.9714^*	0.4413
3	0.6582^*	0.9750^*	0.4460
4	0.6836^*	0.9780^*	0.4409

$^*p < 0.001$

Table 2: Results of the two-way mixed ANOVA that shows the three conditions for which a statistically significant difference was seen in the mean ratings at the third repetition. Description of these speech degradation conditions are as follow: $C07$ stands for "EVS 13.2 kbps SWB", $C15$ is "2 x AMR-WB 6.6kbps" and $C37$ represent "M2M UMTS call AMR-WB 23.85kbps avg. network conditions 3". More details on the conditions can be found at [4].

Condition	$F(3, 53)$	$p - value$	η^2
$C07$	5.23	$= 0.002$	0.093
$C15$	2.891	$= 0.037$	0.054
$C37$	7.287	< 0.001	0.125

3.2 Analysis of Intra-rater Reliability

Additionally, we investigated at which of the repetitions the workers were more confident with their ratings. With this goal in mind, the intra-rater reliability (IRR) was calculated. The IRR, provides a measure of the consistency in the ratings that a single worker gives to the same sample at different points in time. Then, the IRR was computed for each worker considering the first two, three and four repetitions by calculating the intraclass correlation coefficient (ICC) over the ratings of each SQAT test session [23].

The ICC is a statistical method frequently used for assessing IRR for interval, ratio and ordinal variables, and is especially suitable when the "cases" under research are evaluated two or more times [6], like in our study. Since we were interested in the degree of agreement in the absolute values across the ratings from a single worker, we used a "two-way random" model to compute an $ICC(2, 1)$ coefficient [14]. For this purpose, the *"icc"* function of the R package "IRR" was employed using "agreement" and "single" as parameters [6]. The boxplots in Figure 2 show these results. It can be seen that the repetition did not presented any significant effect on the IRR. This also indicates that all of the workers conducted each of the SQAT with high conscientiousness, which in turn confirms further the reliability of the collected ratings.

Moreover, we explored how the Spearman's correlation (ρ) and the RMSE varied through the whole crowdsourcing study (e.g. from repetition one to four). To this end, we computed per worker and per repetition, the correlation and RMSE between the Lab-MOS and the MOS scores given by a single crowd-worker at each of the test session. Figure 3 presents these results with 95% confidence intervals. It can be seen that both the correlation and the RMSE improved

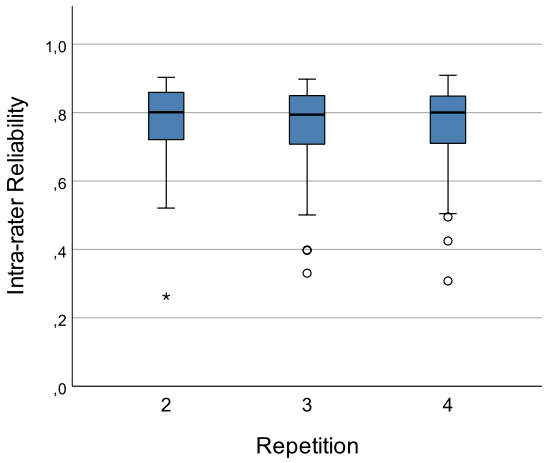


Figure 2: Reliability of the workers at the second, third and fourth time they conducted the SQAT.

with the number of repetitions (e.g. ρ increased from repetition one to four, whereas the RMSE decreased). This outcome evidence that indeed conducting the SQAT multiple times contributed to collect more accurate speech quality scores, and hence better results were achieved after the fourth repetition (when contrasting to the Lab).

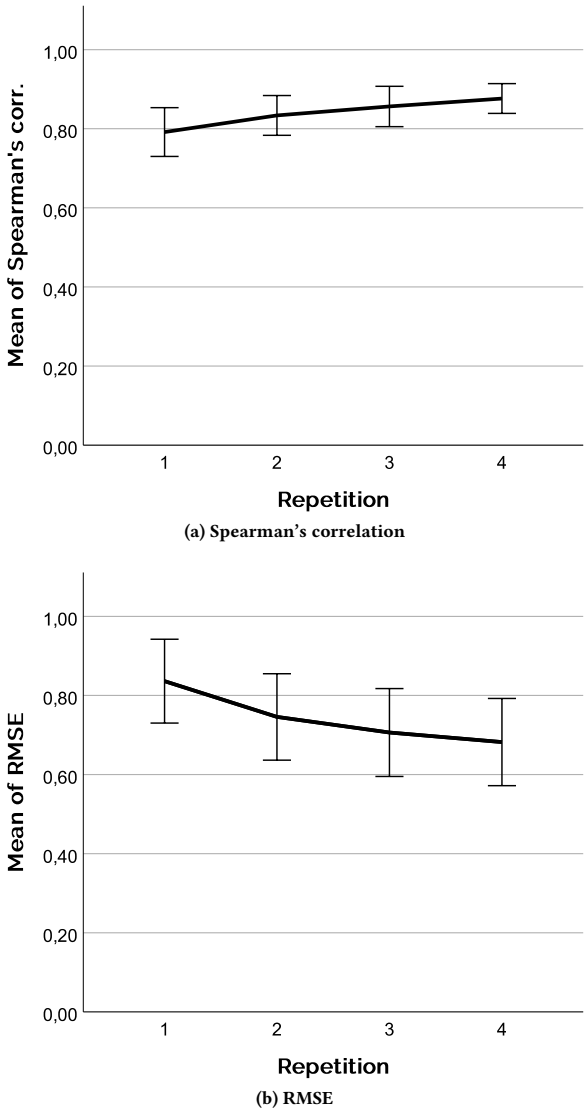
3.3 Discussion

If we remember, our hypothesis was that the inter- and intra-listener agreement would increase gradually from repetition one to four, as well as the accuracy of the speech quality ratings. This later one, in terms of correlation and root-mean-squared-deviation to the Lab results. However, this was not the case with the inter- and intra-rater agreement. While the former fluctuated from the first to the last repetition (see Table 1), the latter remained almost constant as presented in Figure 2. This outcome indicates that there was not a linear relationship between these two metrics, and the fact that listeners as individuals were quite consistent with their answers, it did not contributed to a significant increase in the agreement of the listeners as a group.

Interestingly, an opposite effect was seen with the Spearman's correlation and the RMSE. Both remained almost constant from repetition one to four (see Table 1), when contrasting against the Lab results the ratings of all the workers, whereas both improved with the increase of the repetition number when considering the ratings of individual workers at each of the speech quality assessment test sessions, see Figure 3(a) and 3(b), respectively.

4 CONCLUSION

This work investigate the between and within listeners agreement in a subjective speech quality assessment experiment in crowdsourcing. To this end, a study was conducted in the laboratory in which participants judged the quality of speech files processed with state-of-the-art codecs under different types of transmissions. Ratings were collected in a five-point ACR quality scale in accordance to the ITU-T Rec. P.800 and mean opinion scores were computed. The same study was then replicated in a crowdsourcing platform with



(a) Spearman's correlation

(b) RMSE

Figure 3: Spearman's correlation and root-mean-squared-error (RMSE) with 95% confidence intervals between the ratings provided by each worker and the Lab-MOS.

different participants and a high Spearman's correlation between the Lab-MOS and the CS-MOS was achieved, $r = 0.978$ ($p < .001$).

Our analysis showed that, despite the subjectivity of the task, a proper study setup contributed to collect accurate quality ratings, and the inter-rater agreement increased from the first to the last time the listeners assessed the speech files. However, is worth to point out that a slight decrease in the between listener agreement was seen during the third time the participants conducted the speech quality assessment task. This helped to identify three conditions in the dataset that were quite difficult for the listeners to evaluate. This outcome indicates that disagreement can be also a source of information. Further empirical user studies would be

required to determine the reasons for the differences in the ratings of the aforementioned speech degradation conditions.

ACKNOWLEDGMENTS

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under Grant No.: FKZ 01IS17052.

REFERENCES

[1] 3GPP T S 26.070. [n. d.]. Mandatory speech CODEC speech processing functions; AMR speech Codec; General description.
[2] 3GPP T S 26.171. [n. d.]. Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description.
[3] 3GPP T S 26.441. [n. d.]. Codec for Enhanced Voice Services (EVS); General overview.
[4] Jens Berger and Anna Llagostera. 2018. *A subjective ACR LOT testing super-wideband speech coding in real field measurements and prediction by P.863.* ITU-T Contribution SG12-C.286. International Telecommunication Union, CH-Geneva. 1–11 pages.
[5] Wayne W Daniel. 1980. *Applied nonparametric statistics (2nd ed.).* Boston, MA: Cengage Learning.
[6] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23.
[7] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel. 2014. Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing". https://hal.archives-ouvertes.fr/hal-01078761
[8] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not Enough!. In *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. 131–136. https://doi.org/10.1109/QoMEX.2011.6065690
[9] ITU-T Recommandation P.501. 2017. *Test signals for use in telephonometry.* International Telecommunication Union, Geneva.
[10] ITU-T Recommandation P.800. 1996. *Methods for subjective determination of transmission quality.* International Telecommunication Union, Geneva.
[11] ITU-T Recommandation P.800.1. 2016. *Mean Opinion Score (MOS) Terminology.* International Telecommunication Union, Geneva.
[12] ITU-T Recommandation P.808. 2018. *Subjective evaluation of speech quality with a crowdsourcing approach.* International Telecommunication Union, Geneva.
[13] Maurice George Kendall. 1970. *Rank Correlation Methods* (4th ed.). Charles Griffin.
[14] Richard Landers. 2015. Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS. *The Winnower* (2015). https://doi.org/10.15200/winn.143518.81744
[15] Leonard A Marascuilo and Maryellen McSweeney. 1977. *Nonparametric and distribution-free methods for the social sciences.* Belmont, CA: Wadsworth Publishing Company.
[16] Scott E Maxwell. 1980. Pairwise Multiple Comparisons in Repeated Measures Designs. *Journal of Educational Statistics* 5, 3 (1980), 269–287. https://doi.org/10.3102/10769986005003269
[17] Babak Naderi, Sebastian Möller, and Rafael Zequeira Jiménez. 2018. *Evaluation of the Draft of P.CROWD Recommendation.* ITU-T Contribution SG12-C.204. International Telecommunication Union, CH-Geneva. 1–8 pages. https://www.qu.tu-berlin.de/fileadmin/fg41/publications/naderi{_}2018{_}evaluation-of-the-draft-of-p.crowd-recommendation.pdf
[18] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, and Sebastian Möller. 2015. Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm. In *Interspeech*. ISCA, 2799–2803.
[19] Tim Polzehl, Babak Naderi, Friedemann Köster, and Sebastian Möller. 2015. Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments. In *INTERSPEECH*. 2794–2798.
[20] Alexander Raake. 2007. *Speech Quality of VoIP: Assessment and Prediction.* John Wiley & Sons.
[21] Judith Redi, Ernestasia Siahaan, Pavel Korshunov, Julian Habigt, and Tobias Hoßfeld. 2015. When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal. *Fourth International Workshop on Crowdsourcing for Multimedia* (2015), 33–38. https://doi.org/10.1145/2810188.2810194
[22] Flavio P Ribeiro, Dinei A F Florêncio, Cha Zhang, and Michael L Seltzer. 2011. CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2416–2419. https://doi.org/10.1109/ICASSP.2011.5946971
[23] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 86, 2 (1979), 420.
[24] Kevin P Weinfurt. [n. d.]. *Repeated measures analysis: ANOVA, MANOVA, and HLM. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding MORE multivariate statistics.* 317–361 pages.
[25] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. 2018. Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2018.8463298

TempWeb 2019 Chairs' Welcome

Time is a key dimension to understand the Web. It is fair to say that it has not received yet all the attention it deserves and TempWeb is an attempt to help remedy this situation by putting time as the center of its reflection. Studying time in this context actually covers a large spectrum, from the extraction of temporal information and knowledge, to diachronic studies for the design of infrastructural and experimental settings enabling a proper observation of this dimension.

TempWeb focuses on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Internet scale where the focus is the time dimension. Particular emphasis will be given to temporal Web data analysis that has been collected over extended time periods. A major challenge in this regard is the sheer size of the data it exposes and the ability to make sense of it in a useful and meaningful manner for its users. It is worth noting that this trend of using big data to make inferences is not specific to Web content analytics. Web scale data analytics therefore needs to develop infrastructures and extended analytical tools to make sense of these.

For its ninth edition, TempWeb accepted four very positively reviewed submissions for oral presentation. We interpret the high quality of the submissions and the frequent contributors to TempWeb, as indicators of an evolving community. It shows a clear sign of a positive dynamic in the study of time in the scope of the Web and evidence of the relevance of this effort. In addition, we organized a special session on Web Preservation followed by a panel on this topic that included speakers from The Internet Archive and the Computer History Museum.

We hope that you will find in all these presentations as well as the keynote of Omar Alonso (Microsoft) on "Large Scale Human Sensing Over Time: Challenges and Lessons Learned", plus the discussion and exchanges of this edition of TempWeb, some motivations to look more into this important aspect of the temporal Web.

TempWeb 2019 was jointly organized by University of Caen Normandy (Caen, France), NTENT (Carlsbad, CA, USA), Northeastern University (Silicon Valley, CA, USA), and Internet Memory Research (Paris, France). The workshop proceedings are published by ACM DL as part of The Web Conference 2019 Companion volume.

Marc Spaniol
Chair and Organizer
University of Caen Normandy
France

Ricardo Baeza-Yates
Chair and Organizer
NTENT & Northeastern Univ.
CA, USA

Julien Masanès
Chair and Organizer
Internet Memory Research
France

THE WEB CONFERENCE

TempWeb 2019 Organization

Workshop Chairs & Organizers:
Marc Spaniol *(University of Caen Normandy, France)*
Ricardo Baeza-Yates *(NTENT & Northeastern Univ., USA; UPF, Spain; UChile)*
Julien Masanès *(Internet Memory Research, France)*

Program Committee:
Eytan Adar *(University of Michigan, USA)*
Céline Alec *(University of Caen Normandy, France)*
Omar Alonso *(Microsoft Bing, USA)*
Ralitsa Angelova *(Google, Switzerland)*
Srikanta Bedathur *(IBM Delhi, India)*
Andras A. Benczur *(Hungarian Academy of Science)*
Klaus Berberich *(University of Applied Sciences, Saarbrücken, Germany)*
Ricardo Campos *(INESC TEC / Polytechnique Institute of Tomar, Portugal)*
Renata Galante *(Universidade Federal do Rio Grande do Sul, Brazil)*
Adam Jatowt *(Kyoto University, Japan)*
Nattiya Kanhabua *(NTENT, Spain)*
Scott Kirkpatrick *(Hebrew University Jerusalem, Israel)*
Frank McCown *(Harding University, USA)*
Michael Nelson *(Old Dominion University, USA)*
Kjetil Norvag *(Norwegian University of Science and Technology, Norway)*
Nikos Ntarmos *(University of Glasgow, UK)*
Thomas Risse *(University Library of Frankfurt, Germany)*
Andreas Spitz *(Heidelberg University, Germany)*
Jannik Strötgen *(Bosch Center for Artificial Intelligence, Germany)*
Torsten Suel *(NYU Polytechnic, USA)*
Masashi Toyoda *(Tokyo University, Japan)*
Gerhard Weikum *(Max-Planck-Institut für Informatik, Germany)*

THE WEB CONFERENCE

Large Scale Human Sensing Over Time. Challenges and Lessons Learned

Omar Alonso
Microsoft
omalonso@microsoft.com

ABSTRACT

Twitter and Facebook continue to be top destinations for information consumption on the Internet. The ever-expanding social graph based enables the implementation of traditional features like item recommendation and selection of trending content that rely on human input and other behavioral data. However, given the enormous amount of human sensing in the world at any given moment in any platform, there is a lot of untapped potential that goes beyond simple applications on top of atomic level content like a post or tweet. In this talk we describe a social knowledge graph that discover relationships as they occur over time and how it can be used to capture the evolution of events or stories.

ACM Reference Format:
Omar Alonso. 2019. Large Scale Human Sensing Over Time. Challenges and Lessons Learned . In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3315001

1 INTRODUCTION

Social network posts reflect human sensing at scale. The magnitude of real-time information that is produced can be overwhelming for producers and consumers. If we just look at topics that have high activity volume, there is usually no context for the uninformed user and there is a quite a bit of redundant content for the informed user looking the latest update.

In general, trending topics and hashtags provide a very strong signal for important events that would later be covered in news articles. If the event is of interest to a wider audience, it is highly likely that it would eventually have its own Wikipedia entry. This information flow, from social posts to a final encyclopedia-like entry through a series of news articles is an emerging strategy for producing and consuming content. Current tools are not very well suited for discovering entities, summarizing events, or providing the evolution of a story and related topics. We are interested in mining social network data to construct new data that captures the dynamics and relationships of this large-scale human sensing as we have shown in our prior work [1, 2].

We describe infrastructure that supports querying and retrieving evolving stories about entities and events using a social knowledge graph, called SKG, derived from Twitter. The aim of the system is to fulfill information seeking scenarios by algorithmically generating the core of the story as it evolves over time by using selected relevant content derived from social data and news articles. This new data asset is a new type of document that is not as encyclopedic-centric like a Wikipedia entry but, instead, more dynamic to the many data components of the story, always up-to-date, and constructed in a fashion that allows different aggregations, applications, and, also very important, archiving.

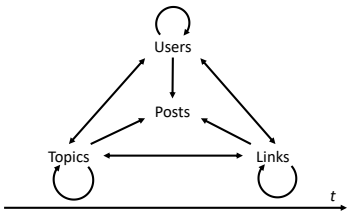


Figure 1: SKG schema: users, topics, links, posts as supporting evidence, and time.

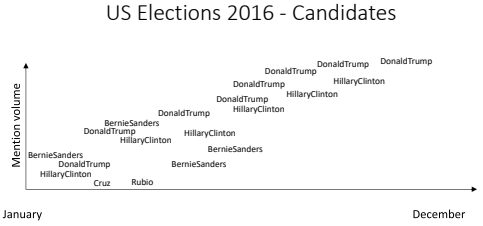


Figure 2: Visualization of the hashtags that represent candidates for the US Elections.

Compared to previous work on knowledge base generation and information extraction from Twitter, we take a bottom-up approach with an emphasis on identifying good quality elements first. Selecting good content, users and links in an efficient manner from the Twitter firehose enables the creation of connections for high quality elements and populate the SKG schema (Figure 1) on a daily basis. The utility of SKG is to retrieve, extract, and present social information as a unit that can be beneficial to many applications.

2 APPLICATIONS

We present a couple of applications. The first one, the evolution of an important event over time using different hashtags and the second, the construction of a Wiki-like document for a specific topic.

In the case of the 2016 US Elections, we can construct a scatter plot of hashtag frequency over time for the main candidates (both

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3315001

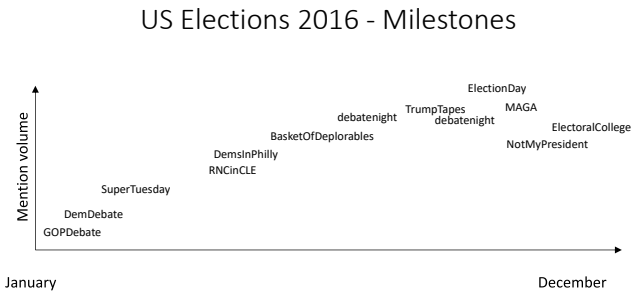


Figure 3: Visualization of the milestones using hashtags that show main subevents (debates, famous phrase) that culminate in the election day in November.

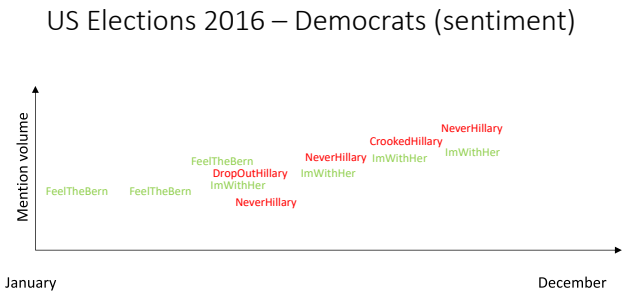


Figure 4: Visualization of the sentiment for Democrats over time. The colors green and red represent positive and negative sentiment respectively.

parties) as presented in Figure 2 for the whole year. We can observe that early in the year there was volume but not as much as we get closer to the elections, which is expected. Many candidates are mentioned in the first quarter of the year and then the two opponents dominate the rest of the chart.

Instead of using candidate names, we now produce a scatter plot of the most significant milestones (Figure 3). From the sequence of pre-established stages in the democratic process (e.g., primaries, debates, election day) to the final outcome (e.g., Electoral College), and the events that occurred within the year like Clinton's famous comment and Trump's tapes, the visualization clearly shows an overview of the elections.

We can also look at sentiment per party over time. For example, in the case of the Democratic party, a scatter plot of positive (`#ImWithHer, #FeelTheBern`) and negative (`#NeverHillary, #CrookedHillary`) hashtags is presented in Figure 4. Similarly, for the Republican party, a scatter plot of positive (`#MAGA`) and negative (`#NeverTrump`) hashtags is presented in Figure 5.

In Figure 6 we show the results of the Wiki-like generation for a topic as a query (`#ces2017`). Say that we are interested in getting an

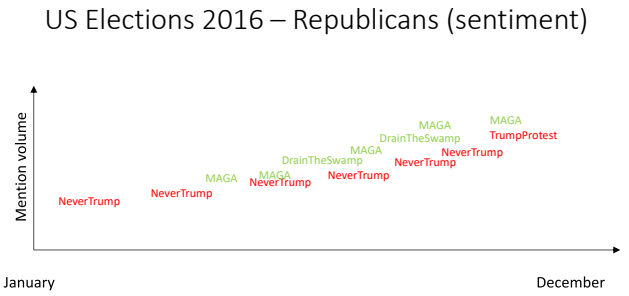


Figure 5: Visualization of the sentiment for Republicans over time. The colors green and red represent positive and negative sentiment respectively.

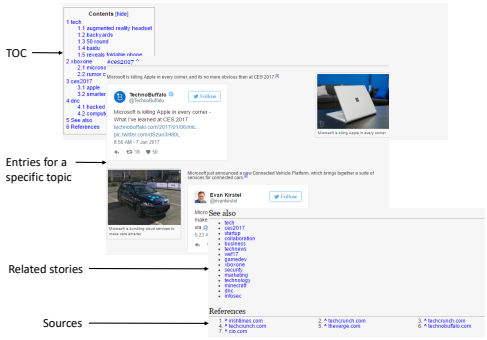


Figure 6: Wiki-like view for the topic `#ces2017`: table of contents, main story, related topics, and sources.

overview of what was discussed at the Consumer Electronics Show in 2017. The system queries SKG and returns a synthetic document that contains the typical elements of a Wikipedia page: table of content, story, related topics, and the sources that were used to create a story. We also include the original tweet as provenance.

3 CONCLUSION

There is little value at the atomic level when reading from a social networking firehose. Instead, we believe that new applications and aggregations can be implemented when looking at social network data as knowledge graphs. In contrast to knowledge-based solutions that retrieve facts, we are interested in stories and related subtopics. We described, at a very high level, a system that supports querying and retrieving evolving stories about events.

REFERENCES

[1] Omar Alonso, Vasileios Kandylas, and Serge-Eric Tremblay. 2018. How it Happened: Discovering and Archiving the Evolution of a Story Using Social Signals. In *Proceedings of JCDL*. 193–202.

[2] Omar Alonso, Vasileios Kandylas, Serge-Eric Tremblay, Jake M. Hofman, and Siddhartha Sen. 2017. What's Happening and What Happened: Searching the Social Web. In *Proceedings of WebSci*. 191–200.

Fast Interval Joins for Temporal SPARQL Queries

Melisachew Wudage Chekol
Data and Web Science Group,
University of Mannheim,
Mannheim, Germany
mel@informatik.uni-mannheim.de

Giuseppe Pirrò
Department of Computer Science,
University of Rome La Sapienza,
Rome, Italy
pirro@icar.cnr.it

Heiner Stuckenschmidt
Data and Web Science Group,
University of Mannheim,
Mannheim, Germany
heiner@informatik.uni-mannheim.de

ABSTRACT

Knowledge graphs enriched with temporal information are becoming more and more common. As an example, the Wikidata KG contains millions of temporal facts associated with validity intervals (i.e., start and end time) covering a variety of domains. While these facts are interesting, computing temporal relations between their intervals allows to discover temporal relations holding between facts (e.g., "football players that get divorced AFTER moving from a team to another"). In this paper we study the problem of computing different kinds of *interval joins* in temporal KGs. In principle, interval joins can be computed by resorting to query languages like SPARQL. However, this language is not optimized for such a task, which makes it hard to answer real-world queries. For instance, the query "find players that were married while being member of a team" times out on Wikidata. We present efficient algorithms to compute interval joins for the main Allen's relations (e.g., BEFORE, AFTER, DURING, MEETS). We also address the problem of interval coalescing, which is used for merging contiguous or overlapping intervals of temporal facts, and propose an efficient algorithm. We integrate our interval joins and coalescing algorithms into a light SPARQL extension called tSPARQL. We evaluated the performance of our algorithms on real-world temporal KGs.

ACM Reference Format:
Melisachew Wudage Chekol, Giuseppe Pirrò, and Heiner Stuckenschmidt. 2019. Fast Interval Joins for Temporal SPARQL Queries. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3314997

1 INTRODUCTION

Knowledge Graphs (KGs) maintaining facts about millions of entities are ubiquitous in many application scenarios, from semantic search [14] to fact checking [7]. Most of existing research in the KG landscape focuses on the analysis of the static part of the data, although KGs like Wikidata contains millions of facts including temporal information that can enrich the available body of knowledge. As an example, a fact like (B. Obama, position, President of USA) carries additional information when considering temporal information, that is, (B. Obama, position, President of USA, 20 January 2009, 20 January 2017). What is also interesting is to go beyond single temporal facts by considering *joins between their validity intervals according to some temporal relation.*

As an example, one can find "who was the German president DURING B. Obama's office", "football players who played in two German teams during OVERLAPPING time periods" or "mayors of a city whose office was one AFTER the other" or "the intervals DURING which a football player was also married". To answer such requests, KGs like Wikidata, DBpedia, and Yago can be queried upon using the SPARQL query language [12]. However, this requires to encode interval comparison using FILTER. On one hand, this will hinder the readability of the temporal part of the query; indeed, a query about OVERLAPS would be more readable if directly using this special keyword. On the other hand, the language is not optimized to answer such queries; indeed, interval comparison based on FILTER will be treated as any other kind of FILTER. To given an example, the query "find players who were married while being a member of a team" times out when executed on the Wikidata SPARQL endpoint[1]. The evaluation of this query could benefit from more efficient ways to solve the interval join problem than using FILTER. For instance, it is well-known that computing the OVERLAPS between sets of intervals does not require pairwise interval comparisons as it can be more efficiently done using Plane Sweep based algorithms [4, 5, 17, 19]. However, these algorithms given two collections of intervals only compute a single kind of interval join, that is, *intersection* (i.e., OVERLAPS). To overcome this issue, the research question that we address in this paper is how to enable the computation of *different kinds of interval joins* in an efficient way in temporal KGs. This is a challenging problem since a pure Plane Sweep-based approach would not allow, for instance, to distinguish between OVERLAPS and DURING, both being different forms of OVERLAPS. On top of that, we are also interested in computing a larger set of temporal relations including BEFORE, STARTS, MEETS, EQUALS and FINISHES.

We also devise an efficient interval coalescing algorithm. Coalescing is the problem of merging overlapping or contiguous intervals of two temporal facts with the same atemporal values. The facts (B. Obama, position, President of USA, 20 January 2009, 20 January 2013) and (B. Obama, position, President of USA, 20 January 2013, 20 January 2017) can be coalesced as (B. Obama, position, President of USA, 20 January 2009, 20 January 2017). We incorporate our algorithms in a light extension of the SPARQL query language called tSPARQL (Interval SPARQL) optimized to compute both interval joins for all Allen's relations [1] and interval coalescing. The previous query about football players that times out when evaluated via SPARQL can be answered in a few seconds via tSPARQL (a comprehensive experimental evaluation will be discussed in Section 5). Moreover,

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314997

[1] http://query.wikidata.org

ID	Temporal Fact
f_1	(Bazoncourt, locatedIn, Moselle, 1790, 1871)
f_2	(Bazoncourt, locatedIn, Bezirk Lothringen, 1871, 1920)
f_3	(Bazoncourt, locatedIn, Moselle, 1920, 2018)
f_4	(Moselle, locatedIn, Grand Est, 2016, 2018)
f_5	(Moselle, locatedIn, Lorraine, 1871, 2015)
f_6	(France, headOfState, Charles DeGaulle, 1959, 1969)
f_7	(France, headOfState, RPoincaré, 1913, 1920)
f_8	(France, containsTerritory, Grand Est, 2016, 2018)
f_9	(France, containsTerritory, Lorraine, 1956, 2015)

Table 1: An excerpt of temporal KG from Wikidata.

iSPARQL allows to express temporal join conditions in a more concise and readable way than the classic SPARQL approach entirely based on FILTER.

Contributions and Outline. We tackle interval join and coalescing problems in large KGs and contribute:

(1) Efficient algorithms to compute different kinds of temporal relations between intervals. Our approach goes beyond approaches that only focus on interval intersection (i.e., OVERLAPS).
(2) An extension of the SPARQL query language called iSPARQL, which allows to write queries in a more succinct way than SPARQL.
(3) An algorithm for coalescing intervals of query answers.
(4) An implementation and experimental evaluation on real-world temporal KGs.

The remainder of the paper is organized as follows. We introduce some background in Section 2. Section 3 introduces iSPARQL. The main algorithms are described in Section 4. In Section 5 we discuss an experimental evaluation. Related Work is reviewed in Section 6. We conclude and sketch future work in Section 7.

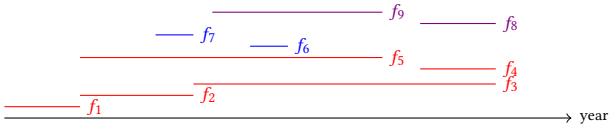
2 PRELIMINARIES

Let I and L be disjoint infinite sets denoting the set of IRIs (identifying resources) and literals (character strings or some other type of data), respectively. We abbreviate the union of these sets ($\mathsf{I} \cup \mathsf{L}$) as IL . We also consider a discrete time domain T as a linearly ordered finite sequence of *time points*; for instance, days, minutes, or milliseconds. A *time interval* is an ordered pair $[t_b, t_e]$ of time points, with $t_b \leq t_e$ and $t_b, t_e \in \mathsf{T}$, which denotes the closed interval from t_b to t_e. We adopt the interval-based temporal domain in our data model. Note that point-based temporal domains can be converted into interval-based domains by using for every time point t an interval $[t, t]$. A quintuple of the form $(\mathsf{s}, \mathsf{p}, \mathsf{o}, t_b, t_e) \in \mathsf{I} \times \mathsf{I} \times \mathsf{IL} \times \mathsf{T}$ is called a *temporal fact*[2]; s is the *subject*, p is the *predicate*, o is the *object*, and t_b and t_e are the start and endpoint of the validity interval $[t_b, t_e]$. The temporal element (interval) represents the time period in which a triple (s, p, o) is valid, i.e., the *valid time* of the triple. A set of quintuples is referred to as a *a temporal knowledge graph*. Table 1 shows an excerpt of temporal KG extracted from Wikidata representing the north eastern region of France. These temporal facts are interesting, for instance, to understand the evolution of regions; one could ask a query to see which regions contained Bezancourt BEFORE Raymond Poincaré came to power or other queries about temporal relations between intervals.

[2] We do not consider blank nodes.

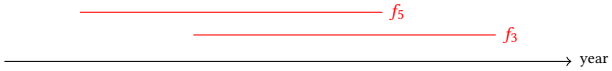
Temporal Relations. We tackle the problem of computing temporal relations between (the validity intervals of) facts. We consider temporal relations defined by Allen [1] and focus on the e following 7 relations plus the inverse of the first 6 (not reported here).

Temporal Relation	Constraint
$[t_b, t_e]$ BEFORE $[t'_b, t'_e]$	$t_e < t'_b$
$[t_b, t_e]$ MEETS $[t'_b, t'_e]$	$t_e = t'_b$
$[t_b, t_e]$ FINISHES $[t'_b, t'_e]$	$t_b > t'_b$ and $t_e = t'_e$
$[t_b, t_e]$ STARTS $[t'_b, t'_e]$	$t_b = t'_b$ and $t_e < t'_e$
$[t_b, t_e]$ DURING $[t'_b, t'_e]$	$t_b > t'_b$ and $t_e < t'_e$
$[t_b, t_e]$ OVERLAP $[t'_b, t'_e]$	$t_b < t'_b < t_e$ and $t_e < t'_e$
$[t_b, t_e]$ EQUALS $[t'_b, t'_e]$	$t_b = t'_b$ and $t_e = t'_e$



(a) Intervals, of the temporal KG in the running Example, on a time line. We can see the evolution of territory changes.

(b) Poincare's presidency (f_7) FINISHES Bazoncourt's location in Bezirk Lothringen (f_2).



(c) Bazoncourt located in Moselle (f_3) OVERLAPS with Moselle located in Lorraine (f_5).

Figure 1: (a) Time intervals of facts ($f_1, \ldots, f_9$) in Table 1, (b) and (c) temporal joins involving FINISHES and OVERLAPS.

Interval Join Problem. Given two collections of intervals R and S and a temporal relation t, the goal is to compute output pairs $(r, s) \in R \times S$ such that the intervals r and s are in the temporal relation t. While existing algorithms mainly focus on the OVERLAPS temporal relation (e.g., [5]) we consider a larger set of relations.

Fig. 1 reports in red some intervals for facts about locatedIn (f_1, f_2, f_3, f_4 and f_5 in Table 1), in violet for headOfState (f_6 and f_7) and in blue for containsTerritory (f_8 and f_9). We can identify the following relations: BEFORE(f_1,f_7), STARTS(f_2, f_7), DURING(f_9, f_3), EQUALS(f_8,f_4) and so on.

Query Language. To query KGs there exists a standard query language called SPARQL [12]. Several extensions have been proposed to handle temporal data (see e.g., [10]). However, none of them

has focused on the problem of computing different kinds of interval joins efficiently. For the most part they focus on syntactic extensions. We will review the most recent proposals in Section 6.

3 ISPARQL: INTERVAL SPARQL

We define our ISPARQL language on top of an extension of the SPARQL query language called SPARQL* [13], which is particularly suitable to query data that use *reification*. SPARQL* leverages a data model called RDF* to concisely represent reified statements (we provide more details in Section 5.1). Let V be a set of variables. An ISPARQL query is a query of the form: SELECT V WHERE {*QP*}. The syntax of SPARQL* query patterns (*QP*) is given below.

$$QP ::= \mathsf{IV} \times \mathsf{IV} \times \mathsf{ILV} \times \mathsf{TV} \times \mathsf{TV} \mid QP_1 \text{ AND } QP_2 \mid \{QP_1\} \text{ UNION } \{QP_2\} \mid \{QP_1\} \text{ MINUS } \{QP_2\} \mid QP_1 \text{ OPT } \{QP_2\} \mid QP \text{ FILTER } (R)$$

Let $x, y \in \mathsf{V}$, $c \in \mathsf{ILT}$ and $t_b, t_e, t'_b, t'_e \in \mathsf{V}$. ISPARQL expresses temporal relations via the SPARQL* built-in expression R formed according to the following grammar:

$$R ::= \text{MEETS}(t_b, t_e, t'_b, t'_e) \mid \text{OVERLAP}(t_b, t_e, t'_b, t'_e) \mid \text{BEFORE}(t_b, t_e, t'_b, t'_e) \mid \text{STARTS}(t_b, t_e, t'_b, t'_e) \mid \text{DURING}(t_b, t_e, t'_b, t'_e) \mid \text{FINISHES}(t_b, t_e, t'_b, t'_e) \mid \text{BEFORE}(t_b, t_e, t'_b, t'_e) \mid !R_1 \mid R_1 \,||\, R_2 \mid R_1 \&\& R_2 \mid \text{BOUND}(x) \mid x = c \mid x < y \mid x! = y$$

The ISPARQL semantics is directly derived from that of SPARQL* [13].

3.1 Running Examples

We now show some examples of ISPARQL. The following temporal query is used to query the KG shown in Table 1.

"Select regions containing Bazoncourt before R. Poincare came to power".

```
SELECT ?x WHERE {
  (Bazoncourt locatedin ?x) ?s ?e.
  (?x locatedin ?y) ?s1 ?e1.
  (?z containsTerritory ?y) ?s2 ?e2.
  (?z headOfState RPoincare) ?s3 ?e3.
  FILTER (OVERLAPS(?s, ?e, ?s1, ?e1)
   && OVERLAPS(?s1, ?e1, ?s2, ?e2)&&
  BEFORE(?s, ?e, ?s3, ?e3) )
 }
```

The above query involves three temporal join operations, namely, two OVERLAPS and one BEFORE. Variable names starting with ?s (resp., ?e) are used to denote start times (resp., end times) of validity intervals. The answer of the query is Moselle which is obtained by computing the temporal joins using OVERLAPS(1790,1871,1871,2015) and OVERLAPS(1871,2015,1956,2015); moreover, the additional test BEFORE(1790,1871,1913,1920) is checked. Another example of an ISPARQL query (the corresponding SPARQL syntax, which is more verbose because reification is needed in order to encode temporal aspects, is shown on the right) is:

"Select athletes that played for some team while being married".[3]

```
SELECT ?x ?y  WHERE {
(?x P54 ?z) ?s1 ?e1.
(?x P26 ?y) ?s2 ?e2.
FILTER
(OVERLAPS(?s1,?e1,?s2,?e2))
}
```

```
SELECT ?x ?y  WHERE {
?x P54 ?reif1. ?x P26 ?reif2.
?reif1 P54 ?z. ?reif2 P26 ?y.
?reif1 P580 ?s1. ?reif1 P582 ?e1.
?reif2 P580 ?s2. ?reif2 P582 ?e2.
FILTER((?s1 <= ?s2 &&  ?s2 <= ?e1)
  ||
(?s2 <= ?s1 &&  ?s1 <= ?e2))
}
```

Evaluating such queries that compute interval joins between sets of intervals can be time consuming on large knowledge graphs such as Wikidata. As an example, the last query *times out* on Wikidata. The objective of this paper is to provide efficient algorithms for computing interval joins over Allen's relations.

[3] P54, P26, P580 and P582 are shorthands for *member of sports team*, *spouse*, *start time* and end time, respectively.

4 INTERVAL JOIN ALGORITHMS

The classical interval join problem, takes as input two collections of intervals R and S, and outputs the pairs $(r, s) \in R \times S$, such that intervals r and s intersect. However, the above examples show that there are scenarios where computing other kinds of relations (e.g., AFTER, DURING) between intervals is interesting. Consider, for example, the time intervals shown in Figure 1(a), if we apply the FINISHES relation, we obtain the intervals (f_7, f_2) shown in Figure 1(b). On the other hand, if we apply the OVERLAPS relation, we obtain the intervals (f_5, f_3). In order to compute such interval joins based on Allen's relations, a naive approach that performs pairwise comparison can be used. However, this approach has a quadratic complexity that limits its applicability in large knowledge graphs. We now discuss efficient algorithms, based on plane-sweep interval join, for the seven Allen relations, namely, BEFORE, MEETS, FINISHES, STARTS, DURING, OVERLAPS, and EQUALS. The algorithms are also applicable to the inverses of these relations.

4.1 OVERLAPS

The interval join based on the OVERLAPS relation for two intervals R and S is defined as: $\text{OVERLAPS}(R, S) = \{(r, s) \mid \forall r = (t_b, t_e) \in R \wedge \forall s = (t'_b, t'_e) \in S \wedge t_b < t'_b \wedge t'_b < t_e \wedge t_e < t'_e\}$. The algorithm for computing the overlap of two intervals is shown in Algorithm 1. The algorithm proceeds as follows: the start and end points of the intervals R and S are maintained in a list L (line 4). Besides, we create four lists that keep track of the active/open and closed intervals (lines 5–6). After sorting L, it is scanned iteratively, for each t in L if t is in R and it is a start point, then it is added to the active set A^R (line 10–13). Otherwise, it is added to the closed set $A^{R'}$ and removed from A^R (line 14–20). On the other hand, if t belongs to S and if it is a start point, then it is added to the active set A^S (line 20–22). Otherwise, it is added to the closed set $A^{S'}$ and removed from A^S (line 23). All those intervals that are in the closed intervals A^R and A^S are intersecting. However, not all of them OVERLAPS. Those that satisfy the OVERLAPS relation are inserted into the output list O (line 28-29). Afterwards, the $A^{R'}$ and $A^{S'}$ are emptied. It is already established that computing interval joins can be performed in linear time [5]. By utilizing an efficient data structure (such as a gapless hash map where insertion, update and deletion can be done in constant time [19]), Algorithm 1 runs in $O(|R| + |S| + K)$ time, where K is the number of results. Our algorithm shares commonalities with the classical interval join based on the plane sweep algorithm [5]. However, the classical interval join does not allow to distinguish between different kinds of interval relations such as OVERLAPS, DURING, MEETSand so on.

Algorithm 1 Plane-Sweep based Interval join for OVERLAPS

```
1:  procedure OVERLAPS(R, S)
2:      Input: lists of intervals R and S
3:      Output: a list O of intervals (r, s) ∈ OVERLAPS(R, S)
4:      a list L ← {s, e, s′, e′ | [s, e] ∈ R, [s′, e′] ∈ S}
5:      A^R ← ∅, A^S ← ∅                    ▷ active intervals from R, S
6:      A^R′ ← ∅, A^S′ ← ∅                  ▷ closed intervals from R, S
7:      O ← ∅
8:      sort L
9:      for t ∈ L do
10:         if t is in R then
11:             [s, e] ← interval in R s.t. t = s or t = e
12:             if t = s is a start point then
13:                 add [s, e] to A^R
14:             else
15:                 if [s, e] ∈ A^R then
16:                     add [s, e] to A^R′
17:                 remove [s, e] from A^R
18:         else
19:             [s, e] ← interval in S s.t. t = s or t = e
20:             if t = s is a start point and A^R ≠ ∅ then
21:                 add [s, e] to A^S
22:             else
23:                 if [s, e] ∈ A^S then
24:                     add [s, e] to A^S′
25:                     remove [s, e] from A^S
26:                     O ← {∀[s, e] ∈ A^R′ ∪ ∀[s′, e′] ∈
27:                         A^S′ | s < s′ ∧ s′ < e ∧ e < e′}
28:                     empty A^S′ and A^R′
29:     return O
```

4.2 Other Temporal Relations

The algorithms for STARTS, FINISHES, DURING and BEFORE are not displayed for the sake of space. However, they are similar to that of OVERLAPS with the main difference lying on the way indexes are created. The MEETS relation can be implemented by indexing (using hash tables). For two sets of intervals R and S, the MEETS relation is defined as $\text{MEETS}(R, S) = \{(r, s) | \forall r = (t_b, t_e) \in R \wedge \forall s = (t'_b, t'_e) \in S \wedge t_e = t'_b\}$. In order to compute MEETS, our algorithm proceeds as follows: we index R by the endpoints and S by start points. Indexing can be done using efficient data structures such as hash tables. Sort the index keys of R. For each key in the index of R, check if the key exits in the index of S. If a key of R is also in S, then the corresponding intervals are involved in a MEETS relation. The algorithm computes the meets joins in $O(|R| + |S|)$ time. Similarly, using hash tables, the EQUALS interval join can be implemented. Like MEETS, the worst time complexity of this operation is $O(|R| + |S|)$.

4.3 Coalescing

Coalescing is the process of merging two facts with the same subject, predicate and object and overlapping or contiguous time intervals. Coalescing can be used to remove duplicate facts. While temporal projection and temporal union can return an uncoalesced answer when evaluated over a coalesced (duplicate free) graph, temporal selection and join operations preserve coalescing when evaluated over a coalesced graph [3]. Consider the following query to select the history of cities and regions from the graph shown in Table 1.

```
SELECT ?x ?s ?e WHERE { (?x locatedin ?y) ?s ?e . }
```

The answer to the above query is shown below. The answers a_1, a_2 and a_3 are uncoalesced. Since these answers are redundant because they contain overlapping intervals, we can apply the coalesce operation in order to remove duplicates.

ID	Uncoalesced answers
a_1	Bazoncourt,1790,1871
a_2	Bazoncourt,1871,1920
a_3	Bazoncourt1920,2018
a_4	Moselle,2016,2018
a_5	Moselle,1871,2015

ID	Coalesced answers
a_1	Bazoncourt,1790,1920
a_4	Moselle,2016,2018
a_5	Moselle,1871,2015

To coalesce query answers, we introduce a new tSPARQL operator called *coalesce()*. This operator is applied on projected temporal variables. As an example, the rewriting of the above query is:

```
SELECT coalesce(?x ?s ?e) WHERE { (?x locatedin ?y) ?s ?e. }
```

It is well known that coalescing is an expensive operation. To the best of our knowledge, there are no algorithms that perform efficient coalescing in SPARQL. Hence, in order to tackle this problem, we propose the procedure shown in Algorithm 2. The algorithm takes as input query answers $A = \{(A_1, I_1), \ldots, (A_n, I_n)\}$ (like the one shown in the above table) of a SPARQL query. The answers are indexed by atemporal values (line 4). For each element in the index (A_i, I_i), its intervals I_i are retrieved and sorted by start point (lines 5–7). And then for each interval $[start, end]$ in I_i, if the list *CoalescedIntervals* is empty, then $[start, end]$ will be added to it lines (8–10). Otherwise, the last entry in *CoalescedIntervals* is retrieved (line 12) and compared with $[start, end]$, if the two intervals intersect (line 14), then the last entry is replaced by the coalesced interval. Otherwise, $[start, end]$ will be added to the coalesced list (line 17). After all the intervals of an answer A_i have been coalesced, we update the index (line 21). Note that this coalesce procedure can be directly applied on the level of query patterns (e.g., COALESCE((?x locatedin ?y) ?start ?end)). The algebra of tSPARQL introduced in Section 3 can be extended as follows:

$$
\begin{aligned}
\text{COALESCE}(QP) ::= \ & \text{COALESCE}(\text{IV} \times \text{IV} \times \text{ILV} \times \text{TV} \times \text{TV}) \mid \\
& \text{COALESCE}(QP_1) \text{ AND } \text{COALESCE}(QP_2) \mid \\
& \{\text{COALESCE}(QP_1)\} \text{ UNION } \{\text{COALESCE}(QP_2)\} \mid \\
& \{\text{COALESCE}(QP_1)\} \text{ MINUS } \{\text{COALESCE}(QP_2)\} \mid \\
& \text{COALESCE}(QP_1) \text{ OPT } \{\text{COALESCE}(QP_2)\} \mid \\
& \text{COALESCE}(QP) \text{ FILTER } (R)
\end{aligned}
$$

The runtime complexity of the algorithm is $O(n \log n)$ where n is the number of distinct answers.

5 IMPLEMENTATION AND EVALUATION

In this section we report on the experimental evaluation of our approach. We describe the datasets and the implementation in Section 5.1. Section 5.2 reports on the performance evaluation of our approach in terms of running time as compared to naive implementation using nested loops. In Section 5.3 we report on the usage of

Algorithm 2 Coalescing query answers

```
1:  procedure COALESCE(A)
2:      Input: query answers A = {(A_1, I_1), ..., (A_n, I_n)}
3:      Output: coalesced answers A' = {(A_1, I'_1), ..., (A_m, I'_m)}
4:      Index ← index answers in A by atemporal values a_i
5:      for each (A_i, I_i) in Index do
6:          CoalescedIntervals ← ∅
7:          sort I_i by start point
8:          for each [start, end] ∈ I_i do
9:              if CoalescedIntervals = ∅ then
10:                 add [start, end] to CoalescedIntervals
11:             else
12:                 [start', end'] ← last entry of CoalescedIntervals
13:                 if max(start, start') ≤ min(end, end') then
14:                     replace last entry of CoalescedIntervals by
15:                     [start', max(end, end')]
16:                 else
17:                     add [start, end] to CoalescedIntervals
18:         Update (A_i, I_i) with (A_i, CoalescedIntervals)
```

iSPARQL to analyze temporal relations between triples in Wikidata expressed using 133 predicates.

5.1 Data Collection

Temporal information is represented in most of existing KGs based on the RDF data model using reification. Reification allows to make statements about statements. In particular, to say that a particular triple f=(s, p, o) is valid in the interval $[t_b, t_e]$ one can use two additional triples, that is, (f, startValidity, t_b) and (f, startValidity, t_e) where f is a statement id. An alternative form of reification (called RDR) has been recently introduced [13] where the above temporal triple can be expressed as <<(s, p, o)>> startValidity t_b; endValidity t_e. The idea is to allow a more concise form of reification. We encoded the dataset used in the experiments according to the RDR syntax. We now report details about the datasets:

- **Wikidata**: Temporal data were collected by first looking at temporal facts, that is, those having P580 (start time) and P582 (end time); in particular, we extracted ~6M temporal facts from 133 temporal relations (e.g., spouse, presidentOf, worksFor, playsFor) .
- **Footballdb**: footballdb.com contains two important relations: *playsFor* and *birthdate*. We extracted ≈20K temporal facts for the *playsFor* relation and >6K facts for the birthdate relation.
- **Synthetic data**: in order to test the scalability of our approach we also generated synthetic data. We created intervals with a normal distribution. In addition, to test the performance of the coalesce algorithm, we designed a synthetic dataset called *non-intersect* in which all the intervals in that dataset are disjoint.

Implementation. We implemented our algorithms in C++ and compiled them using GCC 5.5.0. Besides the algorithms described in Section 4 we also implemented a variant based on grouping (results shown in Figure 2) where the idea is to group consecutively intervals from the same list and produce join results for them in batch, thereby avoiding redundant comparisons. All data used by the algorithms reside in main memory and the coalescing algorithm is single threaded. To call the algorithms from iSPARQL we used special built-in functions. We ran the experiments using the BlazeGraph triplestore, which supports both Reification Done Right (RDR)[4] and built-in calls[5].

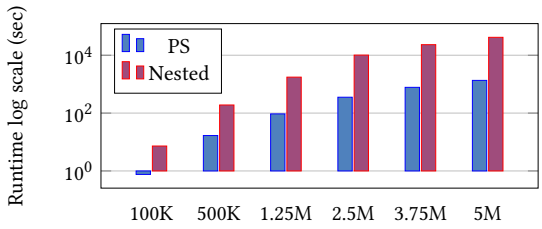


Figure 2: Plain-sweep (PS) vs nested loops on synthetic data.

5.2 Runtime Performance

We compared the running times of our algorithms with that of the standard SPARQL evaluation (using FILTERs) over a synthetic dataset. Results are shown in Figure 2. The runtime performance of plane-sweep (PS) and nested loop (Nested) algorithms on an OVERLAPS query with increasing data size is reported. The PS algorithm is much faster than a naive Nested variant. As can be seen PS is orders of magnitude faster than the naive nested loops. At 5M joins, Nested took 41,610 seconds while PS took just 1,349 seconds.

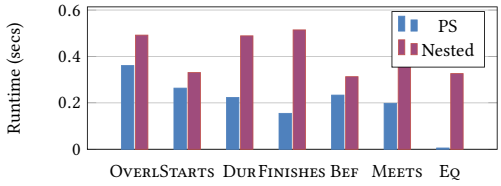
We now discuss the experimental results for some selected queries. Note that our exhaustive analysis is excluded due to space.

- *Query1: find people who were playing for a team while being married.* The iSPARQL syntax of this query is given in the running example. We ran the query on BlazeGraph and performed the joins involving OVERLAPS using Algorithm 1. The runtime performance of the algorithm is shown in Figure 3(a). We compare a nested loop (naive) implementation (*Nested*) of OVERLAPS with that of Algorithm 1 (*PS*). PS outperforms Nested for all of Allen's relations. For this query, the number of EQUALS interval joins (equal intervals) is much smaller than all the other relation as reported in Figure 3(c).
- *Query2: find pairs of people who played for a team during overlapping time periods.* The results of this query are shown in Figure 3(b). As above, PS outperforms Nested for all the relations. We used the same query on the footballdb dataset to count the number of interval join results shown Figure 3(d). As can be seen, the number of OVERLAPS is smaller than all the other relations.

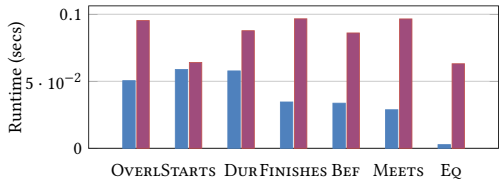
In addition to runtime, we computed the number of joins for each interval relation. For Query1, the join counts are shown in Figure 3(c). The count of the BEFORE relation is larger than all the others. We can interpret the result as most athletes play/work for a team before getting married. We can also see that there are few athletes that end their marriage when they leave a team.

Coalescing Experiment. We tested the performance of our coalescing algorithm on all of our test datasets. The results of our

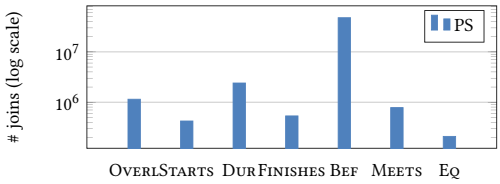
[4] https://wiki.blazegraph.com/wiki/index.php/Reification_Done_Right
[5] https://wiki.blazegraph.com/wiki/index.php/CustomFunction



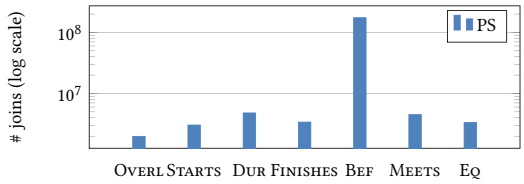
(a) Allen interval joins on Wikidata predicates P54 and P26.



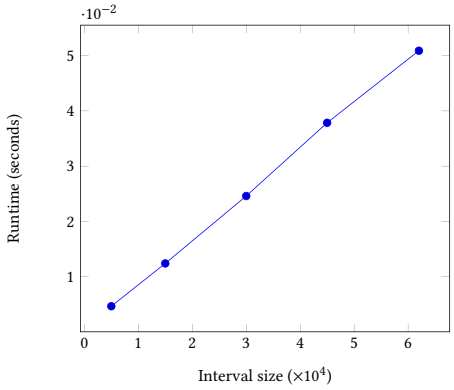
(b) Allen interval joins on Wikidata predicates P54 and P54.



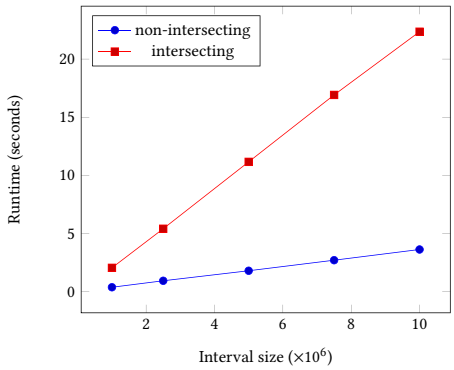
(c) Number of joins on Wikidata predicates P54 and P26.



(d) Number of joins of Footballdb.com plays for relation.



(e) Runtime of COALESCE((?x P54 ?y) ?start ?end) over Wikidata with increasing answer size.



(f) Runtime of coalesce operation on two synthetic datasets (one containing non intersecting intervals).

Figure 3: Runtime comparison of plain sweep and nested loops (a) and (b). The join count for each Allen relation is shown in (c) and (d). Performance of the coalescing algorithm is reported in (e) and (f).

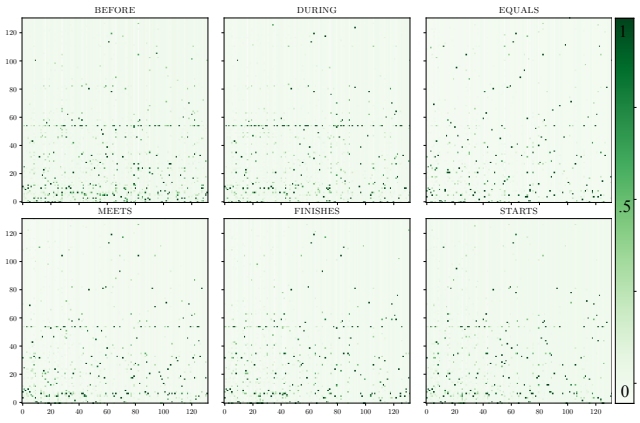


Figure 4: Count of the number of facts expressed via p_i (x-axis) in a given temporal relation with facts expressed via p_j (y-axis). Count values are normalized between 0 and 1.

experiments are shown in Figure 3(e) and (f). When the number of intervals is 5000, coalescing takes less than 4milliseconds (Figure 3(e)) and when the number of intervals is 10million, it takes 3.62 seconds (for non-intersecting) and 22.5 seconds (for intersecting) as shown in Figure 3(f). These results show that our single-scan coalescing algorithm is fast enough to be used for large KGs such as Wikidata and beyond.

5.3 Wikidata Temporal Analysis

tSPARQL is a useful tool to perform temporal-aware analytics on large KGs. As a concrete use-case, we analyzed counts between collections of intervals expressed using 133 Wikidata predicates that concern temporal facts (e.g., `spouse`, `affiliation`). Counts for 6 temporal relations (normalized between 0 and 1) are shown in Fig. 4 where the x and y axes represent predicates; darker colors mean larger counts. The result of this analysis allows to understand, for instance, that facts expressed via the properties `memberOfSportTeam` often (i.e., 312538 times) come BEFORE facts expressed via `headCoach` (e.g., P. Guardiola was playing for Barcelona before becoming the coach), that facts expressed via `workLocation` very often (i.e., 10325 times) occur DURING facts expressed via `positionHeld`, or that facts using the property `workLocation` occur often DURING facts

expressed via `headOfGovernment`. For instance, G. W. Bush lives in Florida now but he lived in Washington while he was president. From this analysis, one can also understand marital tendencies of athletes with respect to team membership.

6 RELATED WORK

Efficiently computing the interval join of two temporal relations has been well studied in temporal databases (see for instance [3–6]). However, this has not been the case for most of the temporal extensions. One prominent example, stSPARQL (aka. strabon) [15] allows temporal joins using Allen's relations, however, the join operations do not use efficient algorithms as done in this study.

The introduction of time into RDF has been studied almost one decade ago [11]. Gutierrez et al. [11] studied fundamental problems of temporal RDF graphs such as entailment and outlined a query language allowing to express queries making usage of intervals. Along these lines several other extensions of SPARQL such as τ-SPARQL [21], T-SPARQL [9], tRDF [20] and RDF-TX [8] have been proposed. τ-SPARQL extends SPARQL query patterns with two variables `?s` and `?e` to bind the start time and end time of temporal RDF triples and express temporal queries. The evaluation is done by rewriting τ-SPARQL queries into standard SPARQL queries. T-SPARQL leverages a multi-temporal RDF model where each RDF triple is annotated with a temporal element that represents a set of temporal intervals. T-SPARQL is based on TSQL2 (temporal SQL). The tRDF system builds upon the Gutierrez et al. [11] temporal RDF model. tRDF queries are evaluated using an index (viz. tGrin) based on a strategy that clusters RDF triples using a graphical-temporal distance. RDF-TX [8] offers both a temporal extension of SPARQL and an indexing system based on compressed multiversion B+ trees. SPARQL-ST is a query language for spatiotemporal RDF data [18]. It extends SPARQL with spatial and temporal variables. The temporal variables appear in the fourth position of valid time temporal triple patterns (i.e., when temporal triples are represented by quads); and thus, these variables can be mapped into time intervals upon query evaluation. Additionally, SPARQL-ST proposed a new filter operator called TEMPORAL FILTER which supports temporal constraints based on Allen's interval relations [1]. Furthermore, an extension of SPARQL-ST called stSPARQL, using the valid time model, is studied in [2, 15], which uses linear constraints to query valid time spatiotemporal RDF data (stRDF). stSPARQL is implemented and integrated into Strabon[6] that extends SPARQL with a set of temporal functions designed based on Allen's interval algebra. It has also functions for time interval intersection, union, and so on. stSPARQL shares the features of ɪSPARQL, the difference lies in the fact that ɪSPARQL computes efficiently using the algorithms proposed in this paper. An approach for representing validity time in RDF(S) and OWL 2 is reported in [16]; authors extend SPARQL by augmenting basic graph patterns with a number of temporal relations such as *during*, *occurs*, *at*, and so on. No implementation is available for the proposed query language. Overall, our goal is to enable the querying of temporal knowledge graphs efficiently.

7 CONCLUSIONS AND FUTURE WORK

We proposed a lightweight extension of SPARQL called ɪSPARQL, which leverages algorithms for efficient computation of interval joins and coalescing. We carried out a number of experiments to showcase the performance of the proposed algorithms. Our findings show that ɪSPARQL is a very good alternative for querying large temporal Web knowledge graphs. Extending ɪSPARQL for spatio-temporal KGs is in our research agenda.

REFERENCES

[1] James F Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM* 26, 11 (1983), 832–843.

[2] Konstantina Bereta, Panayiotis Smeros, and Manolis Koubarakis. 2013. Representation and Querying of Valid Time of Triples in Linked Geospatial Data. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*. 259–274.

[3] Michael H. Böhlen, Richard T. Snodgrass, and Michael D. Soo. 1996. Coalescing in Temporal Databases. In *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India*. 180–191.

[4] Panagiotis Bouros and Nikos Mamoulis. 2017. A forward scan based plane sweep algorithm for parallel interval joins. *Proceedings of the VLDB Endowment* 10, 11 (2017), 1346–1357.

[5] Panagiotis Bouros and Nikos Mamoulis. 2018. Interval Count Semi-Joins. In *Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018*. 425–428.

[6] Bhupesh Chawda, Himanshu Gupta, Sumit Negi, Tanveer A Faruquie, L Venkata Subramaniam, and Mukesh K Mohania. 2014. Processing Interval Joins On Map-Reduce.. In *EDBT*. 463–474.

[7] Valeria Fionda and Giuseppe Pirrò. 2018. Fact Checking via Evidence Patterns.. In *IJCAI*. 3755–3761.

[8] Shi Gao, Jiaqi Gu, and Carlo Zaniolo. 2016. RDF-TX: A Fast, User-Friendly System for Querying the History of RDF Knowledge Bases. In *EDBT*. 269–280.

[9] Fabio Grandi. 2010. T-SPARQL: A TSQL2-like Temporal Query Language for RDF.. In *ADBIS (Local Proceedings)*. Citeseer, 21–30.

[10] Claudio Gutierrez, Carlos Hurtado, and Alejandro Vaisman. 2005. Temporal RDF. In *Proc. of European Semantic Web Conference*. 93–107.

[11] Claudio Gutierrez, Carlos A Hurtado, and Alejandro Vaisman. 2007. Introducing time into RDF. *Knowledge and Data Engineering, IEEE Transactions on* 19, 2 (2007), 207–218.

[12] Steve Harris, Andy Seaborne, and Eric PrudĂćâĆňâĎćhommeaux. 2013. SPARQL 1.1 query language. *W3C recommendation* 21, 10 (2013).

[13] Olaf Hartig. 2017. Foundations of RDF★ and SPARQL★ (An Alternative Approach to Statement-Level Metadata in RDF). In *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017*.

[14] Nandish Jayaram, Arijit Khan, Chengkai Li, Xifeng Yan, and Ramez Elmasri. 2015. Querying knowledge graphs by example entity tuples. *IEEE Transactions on Knowledge and Data Engineering* 27, 10 (2015), 2797–2811.

[15] Manolis Koubarakis and Kostis Kyzirakos. 2010. Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL. In *Extended Semantic Web Conference*. Springer, 425–439.

[16] Boris Motik. 2012. Representing and querying validity time in RDF and OWL: A logic-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web* 12 (2012), 3–21.

[17] Jürg Nievergelt and Franco P. Preparata. 1982. Plane-sweep algorithms for intersecting geometric figures. *Commun. ACM* 25, 10 (1982), 739–747.

[18] Matthew Perry, Prateek Jain, and Amit P Sheth. 2011. Sparql-st: Extending sparql to support spatiotemporal queries. In *Geospatial semantics and the semantic web*. Springer, 61–86.

[19] Danila Piatov, Sven Helmer, and Anton Dignös. 2016. An interval join optimized for modern hardware. In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*. IEEE, 1098–1109.

[20] Andrea Pugliese, Octavian Udrea, and VS Subrahmanian. 2008. Scaling RDF with time. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 605–614.

[21] Jonas Tappolet and Abraham Bernstein. 2009. Applied temporal RDF: Efficient temporal querying of RDF data with SPARQL. In *European Semantic Web Conference*. Springer, 308–322.

[6] Strabon is a spatiotemporal RDF store http://www.strabon.di.uoa.gr

Epitaph or Breaking News? Analyzing and Predicting the Stability of Knowledge Base Properties

Ioannis Dikeoulias
Saarland University
Saarbrücken, Germany
s9iodike@stud.uni-saarland.de

Jannik Strötgen
Bosch Center for Artificial Intelligence
Renningen, Germany
jannik.stroetgen@de.bosch.com

Simon Razniewski
Max Planck Institute for Informatics
Saarbrücken, Germany
srazniew@mpi-inf.mpg.de

ABSTRACT

Knowledge bases (KBs) contain huge amounts of facts about entities, their properties, and relations between them. They are thus the key asset in any intelligent system for tasks such as structured search and question answering. However, due to dynamics in the real world, properties and relations change over time, and stored knowledge may become outdated. While KB information evolves steadily, there is no information whether or not a KB property might be subject to change with high probability or whether it is likely to be stable. Systems exploiting KB information, however, could benefit a lot if they had access to this kind of information. In this paper, we analyze and predict the stability of KB entries, which allows to accompany entries with stability scores. Our predictive model exploits entity-based features and learns through historic data. A particular challenge to determine stability scores is that KB entries are not only added or modified due to real-world changes but also to reduce the incompleteness of KBs in general. Nevertheless, our evaluation of sample properties demonstrates the effectiveness of our method for predicting the one-year stability of KB properties.

KEYWORDS

knowledge bases; temporal validity; stability prediction

ACM Reference Format:
Ioannis Dikeoulias, Jannik Strötgen, and Simon Razniewski. 2019. Epitaph or Breaking News? Analyzing and Predicting the Stability of Knowledge Base Properties. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3314998

1 INTRODUCTION

Motivation. Knowledge bases like Wikidata (WD) are cornerstones of the Semantic Web, and essential for tasks such as entity linking [14], structured search (in particular entity search [3]), and question answering [5]. They usually continuously evolve, by regularly recrawling sources (e.g., DBpedia or YAGO), or by manual editing (e.g., Wikidata), and hence, what is not in them today may be in them tomorrow. Consequently, the question about the extent to which data can be trusted over time evolves. While it has been shown that explicit and implicit temporal information contained in knowledge bases (in combination with temporal tagging of involved texts [15]), can be exploited for entity linking [1], question answering [9], and information retrieval [4], the question about the stability of KB entries has not been raised.

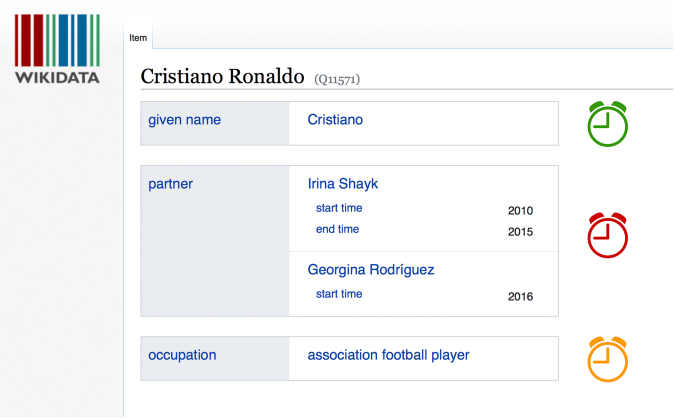


Figure 1: Envisioned visualization of stability information in Wikidata.

Consider for instance Cristiano Ronaldo, 34-year old soccer player currently dating model Georgina Rodriguez. What is the likelihood that the information about his name, occupation, and partner will still be valid in two years?

Contributions. In this paper, we focus on the stability aspect of *subject-property pairs* in Wikidata, like *(Ronaldo, occupation)*. We describe temporal information in Wikidata, characterize and measure the possible changes to subject-property-pairs, and build a predictive model for stability prediction of KB properties. Our evaluation on sample properties of humans shows promising results of our model and improvements over baselines.

The results of our approach are useful for better utilizing KBs in information retrieval settings such as structured search and question answering, but also for NLP tasks (e.g., entity resolution), for understanding KB quality, and for informing and guiding KB editors and readers, as exemplarily shown in Figure 1.

2 BACKGROUND

Related Work. Understanding temporal changes is crucial to estimate KB completeness [13]. A few works have analyzed how to anticipate or manage KB changes, for instance by monitoring relevant news sources [10], by analyzing verb changes in Wikipedia

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314998

articles [18], or by devising update strategies based on a cost-benefit analysis [12]. In [11], the stability of textual descriptions of Wikidata properties was analyzed. An approach based on association rules to predict parts of knowledge bases that are complete at a fixed timepoint is presented in [7]. Approaches to exploit Wikipedia edits for gaining insights about event updates and temporal information retrieval are described in [8] and [17], respectively.

Temporal information in Wikidata. Some properties in WD provide explicit temporal information, for instance *date of birth* and *date of death*. To other properties, temporal information can be added via so-called qualifiers, most importantly *point in time*, *start time* and *end time*, allowing to express, for instance, that Rodriguez is Ronaldo's partner since 2016 (see Fig. 1). Yet, in general, these are scarce, existing only for around 2-3% of statements.[1] A third source of temporal information is the WD revision history, which record the timepoint of the creation and modification of each statement.

3 ANALYZING KB MODIFICATIONS

For our analysis, we compared two Wikidata dumps from January 2017 and January 2018. We then identified the cases where subject-predicate pairs observed a change in frequency. In this work, the focus lies on the analysis of properties about humans, finding, for instance, for human subjects and the property *child*, a total of 21,328 changes. From all changes for humans, 89% were cases where objects for a property were newly created, 7% were cases where additional objects were added, 3% were deletions of all objects, and 1% were reductions.

Types of Property Changes. To get a better understanding of the types of changes of the properties about humans, we analyzed the 20 most common properties regarding whether they were newly added or subject to an increase.

As shown in Figure 2, *occupation* (+2.2%), *award received* (+1.2%), and *educated at* (+0.7%) were those observing the most increases (i.e., additional information was added). In contrast, *gender* (+11.8%), *occupation* (+9.8%), and *date of birth* (+8.5%) were most frequently newly created. We also identified a positive correlation of changes with the number of existing properties per entity, and a bell-shaped relation with the number of existing values per predicate (peak=5 values).

Reasons for Property Changes. Intuitively, the age of a person plays a big role regarding the likelihood that a property of a human changes. However, the reason for changes of KB properties can be two-fold: 1) real-world changes ("breaking news"), and 2) completion of old facts ("epitaph"). In the following, we discuss one Wikidata property with explicit time information (*child*) and one property with time information via qualifiers (*academic degree*).

Figure 3 shows the frequency of real-world changes that are reflected in the KB within the one year of our analysis, versus the changes in the KB that correspond to facts that happened earlier in reality grouped based on the age of a human. As one can see, the two sets of curves are very different. Changes in the real world happen predominantly between the age of 20 and 50, whereas change rates in the KB steadily increase with the age of persons, peaking at an age of 90, and only gradually decreasing beyond that age.

[1] According to our analysis of Wikidata dumps from 2017 and 2018.

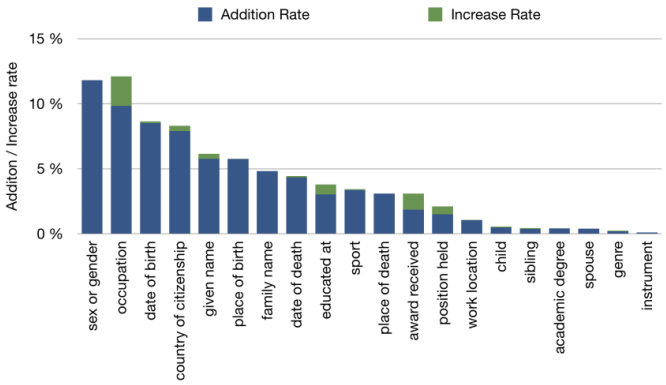


Figure 2: Addition and change rate of most frequent human properties.

This highlights that KBs predominantly lag behind with the real-world, and add data much later in life, e.g., only once persons become famous, or even upon death. This observation makes it particularly challenging to predict changes of knowledge base facts as they do not only rely on changes in the real world – which would often already be challenging to predict [2].

4 PREDICTING STABILITY OF KB PROPERTIES

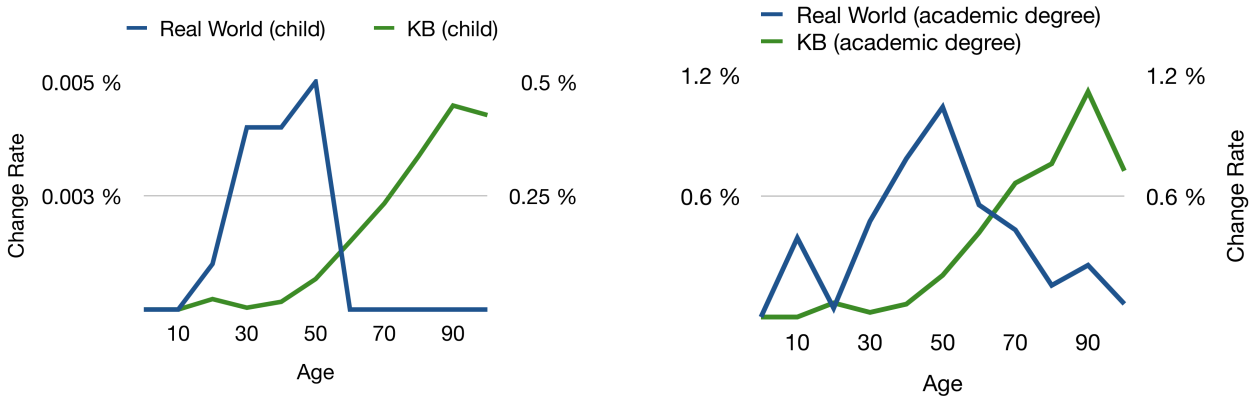
The second main goal of our work is to predict stability in knowledge bases. In particular, we looked at the following problem.

Problem 1 (Stability prediction). *Let e be an entity and p be a property. Given the current timepoint t and a future timepoint t', what is the probability that a fact in the set of all p-facts for e remains unchanged between t and t'?*

An instance of the problem is shown in the introduction: Assuming today is March 6, 2019, what is the likelihood that the *partner* information for Ronaldo will be unchanged on March 6, 2019? Note that for this problem, we can naturally only use information available as of March 6, 2018, and cannot continuously monitor textual sources, as done in [18] and [10].

Approach. What we can use is current information about the entity. In particular, we use the presence of properties and values. We proceed in three steps, the construction of feature vectors, model training, and model application.

(1) *Build entity-level feature vectors.* We construct binary feature vectors for entities by introducing 43 binary features that indicate for each of the 43 most common properties their presence or absence at time t. We also introduce binary features that, for each of the 43 properties, indicate the presence or absence of the 5 most frequent properties, if these occur at least 50 times. The rational is that for properties such as occupation, frequent values such as *politician* and *scientist* are possibly informative, whereas for properties such as child, entities rarely share objects at all. In the end, we thus introduce 149 additional features for object values.



(a) Child.

(b) Academic degree.

Figure 3: Real-time (left scale) vs. late (right scale) additions for two properties.

(2) *Train models based on historic data.* We next train predictive models using two historical Wikidata snapshots. Our target variables are again binary, one variable per property, indicating whether the respective property has changed between the two timestamps. In particular, we use logistic regression as model, because regression allows to learn probabilities even though given only binary values for the target variables.
(3) *Apply the models to individual entities.* Once having trained the models, we can apply them to predict future change probabilities. We feed a feature vector for a given entity to the regression model, and obtain as output a stability probability.

Experimental Setup. We verified the feasibility of this approach using the Wikidata dumps from January 2017 and 2018, on three properties, *child*, *academic degree* and *occupation*. In each case, we trained with 60% of the 3.7 million human entities in the 2017 dump, and used the remaining 40% for testing. A challenge for regression were correlations in our feature set, thus, we scaled down our feature set based on correlation analysis to just the 10 most distinctive features per target variable.

Baselines. We consider three baselines, namely to predict that exactly those persons that are *alive*, *dead*, or *active* (between 20 and 60 years old) are the ones that are going to observe change. The first baseline mirrors the expectation that only alive entities change and do so sufficiently often, while the second quantifies the reverse assumption, that entries of dead subjects are the focus of changes. The third represents a further refinement of the first, focusing on the typical age in which people are active.

Results. To systematically evaluate the performance of our approach, we translate the output probabilities back into binary *change/no change* predictions, and report precision, recall, and f-score in terms of the *change* class. Note that accuracy is not a useful measure, because, due to the high class imbalance, even classifiers that would entirely predict *no change* would achieve >98% accuracy. This becomes also evident via the baselines, which all obtain very low precision scores, due to predicting far too many changes. Furthermore, also their recall shows no clear preference between *alive* and *dead*, illustrating that changes are not confined to one of these sets.

Table 1: Change prediction results of the baselines (BL) and our prediction model (model).

	child			degree			occupation		
	P	R	F	P	R	F	P	R	F
BL-Alive	.00	.23	.00	.00	**.55**	.01	.13	**.71**	.22
BL-Dead	.00	**.77**	.01	.00	.45	.00	.04	.29	.07
BL-Active	.00	.03	.00	.00	.12	.00	.09	.20	.12
Model	**.05**	.12	**.07**	**.87**	.42	**.57**	**.51**	.43	**.47**

Our regression models, in contrast, achieved 5%, 87% and 51% precision and 12%, 42%, and 43% recall for the properties child, degree, and occupation, respectively. For degree and occupation, the high precision values are impressive when taking into account the rather low number of changes overall. In addition, the recall values are still on a promising level for these two properties.

Note that predicting future changes based on present states is truly difficult, because none of the individual features gives a truly strong indication for an upcoming change, and the class imbalance makes the prediction even harder. The child property seems to be particularly challenging probably because child information is often modified in Wikidata once a child becomes famous and thus a Wikidata entity.

Analysis. We can also inspect individual weights learned by the regression, finding for instance that the existence of children is a strong indicator for further changes in the *child* predicate (weight +9.0), while the existence of a father or siblings have a smaller negative influence (-0.5/-0.4). Interestingly, even the existence of *place of death* and *place of burial* have small positive weights w.r.t. change (+0.05/+0.13).

5 CONCLUSIONS AND ONGOING WORK

Our work is a first step towards understanding and predicting stability in knowledge bases. Interesting findings so far are that real-world and KB changes can occur very much disconnected, and that our model achieves promising results for the challenging task of stability prediction. Nevertheless, we are planning to extend this work along several dimensions:

(1) Features: Extending the idea from [18], we aim to include textual features, with the hope that these could indicate not only past, but also future change, and latent representations via knowledge graph embeddings [16], as well as features derived from temporal signatures for entites, as introduced in [1] for time-aware named entity disambiguation.
(2) Models: We plan to utilize other models, in particular, interpretable decision trees, and performant neural models.
(3) Change interrelation: We are currently working on analyzing the circumstances when these changes are connected or disconnected, in particular, whether certain changes co-occur, or follow particular order sequences, similar to the currency constraints in [6].

Furthermore, our goal is to deploy a prototype like shown in Figure 1 as plugin in Wikidata.

REFERENCES

[1] Prabal Agarwal, Jannik Strötgen, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-Aware Named Entity Disambiguation for Diachronic Corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL '18)*. 686–693. http://aclweb.org/anthology/P18-2109

[2] Ricardo Baeza-Yates. 2005. Searching the Future. In *Proceedings of the Mathematical/Formal Methods in Information Retrieval Workshop at SIGIR*.

[3] Krisztian Balog, Marc Bron, and Maarten De Rijke. 2011. Query Modeling for Entity Search Based on Terms, Categories, and Examples. *ACM Transactions on Information Systems (TOIS)* 29, 4, Article 22 (Dec. 2011), 22:1–22:31 pages. https://doi.org/10.1145/2037661.2037667

[4] Leon Derczynski, Jannik Strötgen, Ricardo Campos, and Omar Alonso. 2015. Time and information retrieval: Introduction to the special issue. *Information Processing & Management* 51, 6 (Nov 2015), 786 – 790. https://doi.org/10.1016/j.ipm.2015.05.002

[5] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open Question Answering over Curated and Extracted Knowledge Bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1156–1165. https://doi.org/10.1145/2623330.2623677

[6] Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. 2013. Inferring Data Currency and Consistency for Conflict Resolution. In *2013 IEEE 29th International Conference on Data Engineering (ICDE '13)*. 470–481. https://doi.org/10.1109/ICDE.2013.6544848

[7] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M. Suchanek. 2017. Predicting Completeness in Knowledge Bases. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, 375–383. https://doi.org/10.1145/3018661.3018739

[8] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. 2013. Extracting Event-Related Information from Article Updates in Wikipedia. In *35th European Conference on Information Retrieval (ECIR '13)*. 254–266. https://doi.org/10.1007/978-3-642-36973-5_22

[9] Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, 1807–1810. https://doi.org/10.1145/3269206.3269247

[10] Jiaqing Liang, Sheng Zhang, and Yanghua Xiao. 2017. How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. 3749–3755. https://doi.org/10.24963/ijcai.2017/524

[11] Thomas Pellissier Tanon and Lucie-Aimée Kaffee. 2018. Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. 1801–1803. https://doi.org/10.1145/3184558.3191643

[12] Simon Razniewski. 2016. Optimizing Update Frequencies for Decaying Information. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. 1191–1200. https://doi.org/10.1145/2983323.2983719

[13] Simon Razniewski and Gerhard Weikum. 2018. Knowledge Base Recall: Detecting and Resolving the Unknown Unknowns. *ACL SIGWEB Newsletter* Spring, Article 3 (June 2018), 3:1–3:9 pages. https://doi.org/10.1145/3210578.3210581

[14] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (Feb 2015), 443–460. https://doi.org/10.1109/TKDE.2014.2327028

[15] Jannik Strötgen and Michael Gertz. 2016. *Domain-sensitive Temporal Tagging*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool.

[16] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14)*. 1112–1119. http://dl.acm.org/citation.cfm?id=2893873.2894046

[17] Stewart Whiting, Joemon Jose, and Omar Alonso. 2014. Wikipedia as a Time Machine. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. 857–862. https://doi.org/10.1145/2567948.2579048

[18] Derry Tanti Wijaya, Ndapandula Nakashole, and Tom Mitchell. 2015. "A Spousal Relation Begins with a Deletion of engage and Ends with an Addition of divorce": Learning State Changing Verbs from Wikipedia Revision History. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP '15)*. 518–523. https://doi.org/10.18653/v1/D15-1059

The Power of Temporal Features for Classifying News Articles

Lukas Lange
Saarland University
Saarbrücken, Germany
llange@lsv.uni-saarland.de

Omar Alonso
Microsoft
Mountain View, CA, USA
omalonso@microsoft.com

Jannik Strötgen
Bosch Center for Artificial Intelligence
Renningen, Germany
jannik.stroetgen@de.bosch.com

ABSTRACT

Temporal information extracted from texts and normalized to some standard format has been exploited in a variety of tasks such as information retrieval and question answering. Classifying documents into categories using temporal features has not yet been tried. Such a method might be particularly valuable when classifying sensitive texts such as patient records, i.e., whenever the pure content of the documents should not be used for the classification. In this paper, we describe, as a proof-of-concept, our work on classifying news articles exploiting only features defined over extracted and normalized temporal expressions. Our evaluation of two classification models on large German and English news archives shows promising results and demonstrates the discriminative power of temporal features for topically classifying text documents.

KEYWORDS

temporal information; news classification; temporal n-grams

ACM Reference Format:
Lukas Lange, Omar Alonso, and Jannik Strötgen. 2019. The Power of Temporal Features for Classifying News Articles. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3315000

1 INTRODUCTION

Motivation. Classifying news articles according to a set of categories depending on the articles' content is a standard task. The terms in the texts can be used as weighted features to precisely classify a document as belonging to a particular category such as *politics* or *sports*, e.g., using SVMs [6]. More sophisticated methods than bag-of-words models can be applied, e.g., by using n-grams.

However, in some scenarios, classification needs to be performed without looking deeply into the texts' content, e.g., due to privacy issues when classifying emails or patient records. One approach to avoid using content terms is to apply character-level methods, e.g., character-level convolutional networks [10]. An alternative is to use previously extracted and normalized pieces of information. A particular type of such information is temporal information.

Related Work on Exploiting Temporal Information. With the availability of temporal taggers to extract and normalize temporal expressions from texts [9], more and more applications are exploiting temporal information extracted from text documents. There is little work on using temporal expressions for classifying news articles like our proposed technique. Rocha et al. [7] investigate techniques for context selection and provide an algorithm that identifies which temporal context is best for a classification task. Chambers proposes learning models that use temporal expressions for labeling documents with a correct timestamp [4].

In [3], temporal expressions in documents of a diachronic news archive have been used to analyze how the past is remembered in different countries. In temporal IR, temporal information can be considered as a query topic to specify temporal constraints [1]. The probably most related work to our approach is [2], where search results have been clustered on a timeline, i.e., temporal information was used for classifying documents, however, along timelines and not according to topical categories as in our work.

Contributions. We present our approach to exploit temporal information extracted from texts and normalized to a standard format for topically classifying news articles. Besides the curiosity to determine the discriminative power of temporal features for text classification, the goal of our work is to validate how well texts can be classified only using temporal information extracted from respective documents as an alternative to other methods.

2 METHODS

We use solely temporal expressions and their characteristics to define features for our classification approach.

Temporal Expressions. TimeML (www.timeml.org) defines the `TIMEX3` tag to annotate date (e.g., today), time (e.g., 9 pm), duration (e.g., two days), and set (e.g., weekly) expressions. Its *value* attribute contains normalized information, e.g., about the length of a duration or the anchoring of a date on a timeline. Dates and times can be of different granularities and are realized explicitly (May 2018), implicitly (Black Friday 2018), relatively (next month) or underspecified (April) [9] and refer to the past, present, or future.

Temporal Classification Features. In the following, we describe all features we use for classifying news articles. All features are calculated per document. As basic features (*Types* in Sec. 3), we use

- the numbers of `TIMEX3` annotations per `TIMEX3` type, i.e., number of date, time, duration, and set expressions

For more advanced features, we exploit the *value* attribute of `TIMEX3` annotations. However, fully normalized *value* information, e.g., `2016-11-23` for "November 23, 2016", is very sparse across a document collection. Thus, we generalize the normalized information, e.g., `2016-11-23` as Day and `2018-04` as Month. Based on these generalized *values*, we define the *Value* and Bigrams features:

- a counter for each granularity (e.g., Day, Week, Month, ...)
- temporal bigrams (2 consecutive temporal expressions)

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3315000

Table 1: Statistics on New York Times and Die Zeit corpora.

NYT	# Docs	# Timex	# Dates	# Times	# Durations	# Sets
Arts	88,822	1,258,655	947,222	134,459	159,568	18,406
Business	113,305	593,669	551,272	550	41,043	804
Sports	72,004	1,114,058	816,976	88,499	193,300	15,283
Politics	82,489	1,183,698	944,445	50,275	177,160	11,818
Die Zeit	# Docs	# Timex	# Dates	# Times	# Durations	# Sets
Arts	32,099	296,198	244,967	4,100	44,350	2,781
Business	20,572	222,577	168,215	2,023	46,420	5,919
Sports	830	6,637	4,474	204	1,420	139
Politics	31,575	281,799	216,445	4,284	56,652	4,418

For instance, a YEAR-granularity expression followed by one with DAY-granularity form the temporal bigram [YEAR, DAY] that is found in: "*This year*, the championships started on the *6th of May*."

Further advanced features (referred to as *Relations* in Sec. 3) are the reference direction of a temporal expression (past, present, future) with respect to the document's publication date. That is, using the above features we replaced, for instance, DAY with DAY-FUTURE for the expression *tomorrow* and YEAR with YEAR-PRESENT for *this year*. In addition, we use the realization type of the temporal expressions, i.e., assigned information if an expression occurred explicitly or relatively etc. For instance, *2001* (assuming a document publication date of 2001) would be represented as YEAR-PRESENT-EXPLICIT, while *tomorrow* as DAY-FUTURE-RELATIVE.

Data and Preprocessing. We use the *New York Times* corpus, a collection of English daily news articles from 1987 to 2005 and the *Die Zeit* corpus which contains German articles of a weekly newspaper from 1995 to 2011. The documents from both corpora were automatically annotated with the temporal tagger HeidelTime [8].

For our experiments, we selected the newspaper categories Arts (literature and music), Business (companies), Sports (football and baseball), and Politics (elections and terrorism). The data sets were split randomly into test and training sets, with a test set size of 1,000 documents per category. All remaining documents were used for training. Table 1 provides statistics about the corpora. Note that the category information was extracted from the documents' meta data. While all categories of the English data set contain about 70,000 to 110,000 documents, the German data set has only very few sports articles, which we thus excluded in our experiments.

Classification. We use WEKA [5] to train k-Nearest-Neighbours and Decision Trees with the above described features.

3 EVALUATION

A simple baseline to classify documents is to assign the same class to all documents. With fixed test sizes this results in a 25% success rate for 4 classes. As shown in Table 2(a), rather simple features like the counts of the 4 TIMEX types give far better results with up to 51.5%. A reason is probably that the distribution of expression types is quite different across categories, e.g., art documents contain many TIME-typed expressions, but only few occur in business reports.

The generalized normalized values and the temporal bigrams based on them give even better results with up to 66.3% and 64.7%, respectively. Even if the bigrams seem a little less descriptive for the classification, combining both features leads to a 66.7% success rate. The best results were achieved with the generalized normalized values extended with their temporal relation to the publication date and their realization (decision tree: 69.3%, 9-NN: 68.2%).

Table 2: Accuracy in % with 4 classes (a) and 3 classes (w/o Politics class) (b) for the New York Times corpus, and 3 classes for the Die Zeit corpus (w/o Sports class) (c).

(a)	Types	Values (V)	Bigrams	V + Bigrams	V + Relations
9-NN	51.3	64.5	63.0	66.7	68.2
Decision Tree	51.5	66.3	64.5	65.6	**69.3**
(b)	Types	Values (V)	Bigrams	V + Bigrams	V + Relations
9-NN	64.6	78.7	77.1	77.1	82.3
Decision Tree	65.7	80.8	80.1	80.4	**83.3**
(c)	Types	Values (V)	Bigrams	V + Bigrams	V + Relations
9-NN	41.4	47.3	46.0	46.0	50.4
Decision Tree	42.0	47.0	46.8	47.1	**51.4**

The decision tree classifier also outperforms the 9-NN classifier for tests with 3 classes with 83.3% for New York Times (Table 2(b)) and 51.4% for Die Zeit (Table 2(c)). However, classification of the German texts is not as good compared to their English counterparts, even though there is a similar increase for the different features. This could be explained by the fact that the New York Times training sets were three to four times larger. In addition, the weekly newspaper Die Zeit has longer and more unique-style documents compared to the daily New York Times articles, which are more standard and seem to have more prototypical temporal signatures.

4 SUMMARY AND ONGOING WORK

We presented the first experiments towards classifying documents solely based on temporal information extracted from respective texts. We have developed time-centric features and tested classifiers that achieve up to 83.3% accuracy for 3 classes and 69.3% for 4 classes. Though the results are encouraging, there is room for improvements. We currently test our approach using more classes. We plan to also switch the domain of the documents to test if similarly promising results can be achieved on privacy-sensitive data.

REFERENCES

[1] Prabal Agarwal and Jannik Strötgen. 2017. Tiwiki: Searching Wikipedia with Temporal Constraints. In *Proc. of WWW (WWW'17 Companion)*. 1595–1600. https://doi.org/10.1145/3041021.3051112

[2] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2009. Clustering and Exploring Search Results Using Timeline Constructions. In *Proc. of CIKM*. 97–106. https://doi.org/10.1145/1645953.1645968

[3] Ching-man Au Yeung and Adam Jatowt. 2011. Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. In *Proc. of CIKM*. 1231–1240. https://doi.org/10.1145/2063576.2063755

[4] Nathanael Chambers. 2012. Labeling Documents with Timestamps: Learning from their Time Expressions. In *Proc. of ACL*. 98–106. http://dl.acm.org/citation.cfm?id=2390524.2390539

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[6] Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proc. of ECML*. 137–142. http://dl.acm.org/citation.cfm?id=645326.649721

[7] Leonardo Rocha, Fernando Mourão, Adriano Pereira, Marcos André Gonçalves, and Wagner Meira, Jr. 2008. Exploiting Temporal Contexts in Text Classification. In *Proc. of CIKM*. 243–252. https://doi.org/10.1145/1458082.1458117

[8] Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation* 47, 2 (2013), 269–298. https://doi.org/10.1007/s10579-012-9179-y

[9] Jannik Strötgen and Michael Gertz. 2016. Domain-Sensitive Temporal Tagging. *Synthesis Lectures on Human Language Technologies* 9, 3 (2016). https://doi.org/10.2200/S00721ED1V01Y201606HLT036

[10] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Proc. of NIPS*. 649–657. http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf

Exploring Video Game Searches on the Web

Behrooz Mansouri

B. Thomas Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA, bm3302@rit.edu

Mohammad Sadegh Zahedi

School of Electrical and Computer engineering, University of Tehran, Tehran, Iran, s.zahedi@ut.ac.ir

Ricardo Campos

Polytechnic Institute of Tomar - Smart Cities Research Center, Tomar, Portugal, ricardo.campos@ipt.pt
LIAAD – INESC TEC, Porto, Portugal

Mojgan Farhoodi

Information Technology Faculty, Iran Telecommunication Research Center, Tehran, Iran, farhoodi@itrc.ac.ir

ABSTRACT

As video games are developing fast, many users issue queries related to video games in a daily fashion. While there were a few attempts to understand their behavior, little is known on how the video game-related searches are done. Digesting and analyzing this search behavior may thus be faced as an important contribution for search engines to provide better results and search services for their users. To overcome this lack of knowledge and to gain more insight into how video game searches are done, we analyze in this paper, a number of game search queries submitted to a general search engine named Parsijoo. The analysis conducted was performed on top of 372,508 game search records extracted from the query logs within 253,516 different search sessions. Different aspects of video game searches are studied, including, their temporal distribution, game version specification, popular game categories, popular game platforms, game search sessions and clicked pages. Overall, the experimental analysis on video game searches shows that the current retrieval methods used by traditional search engines cannot be applied for game searches, thus, different retrieval and search services should be considered for these searches in the future.

CCS CONCEPTS

• Information systems ~Web log analysis

KEYWORDS

Video game search, Query log analysis, Web search

1 Introduction

Great advances in video games, made them one major target for spending leisure time for people of different ages. Based on the report from Global games market intelligence firm Newzoo[1] in 2017, there are 2.2 billion active gamers in the world, of which 47%, or 1.0 billion gamers, spend money while playing. Video games became quite popular in the 1980s when arcade video games, gaming consoles and home computer games were introduced to the general public. More recently, the emergence and the spread of cellphones with high speed internet connection together with devices dedicated to games, such as PlayStation4, Xbox and online dissemination app platforms such as Google Play, and Apple Store has created a new trend dedicated to online games.

Currently, there are many websites available providing information regarding video games for their users. For instance, Ign (Imagine Games Network) (www.ign.com) is a video game and entertainment media

[1] https://newzoo.com/insights/articles/newzoo-2017-report-insights-into-the-108-9-billion-global-games-market

website providing information on video games for different consoles such as Xbox, PlayStation and PC. Another game-oriented website is GameFAQs (www.gamefaqs.gamespot.com), which hosts FAQs and walkthroughs for video games. The site has a database of video game information, cheat codes, reviews, game saves, box art images and screenshots, almost all of which is submitted by volunteer contributors. Recently, there has been some effort to develop video game-oriented search engine. For instance, Piwag (www.piwag.com), was founded in March 2016 as a search engine for video games where also gamers can share their experience from around the world. However, to build a vertical search engine dedicated to video games, gamers' search behavior should be studied first. Despite the existence of several websites related to video games, little is known on how game searches are done and to the best of our knowledge, this paper is the first to investigate this matter. We try to shed light on different aspects of gamers search behavior that can be used by search engines to provide better search services to their users. For instance, looking at the temporal distribution of video games searches, may give us information on the days or on the period of the day, where people have more leisure time. One such information can have multiple advantages for search engines and content providers. For example, game advertisement which is a quite common way of getting gamers attention, may be put online at specific days of the week. Also, general search engines can provide query suggestion and query auto-completion for video games at specific times of the day, especially for users interested in video games (based on their search profile) as the chance of a user looking for video game is higher. Another observation is the correlation between the devices that the query was issued from and the video game platform they were looking for. Mobile users are mostly looking for mobile games while queries issued from PC, are particularly tuned to PC video games. Search engines should consider this point while retrieving the documents, giving higher ranks to webpages related to the device from which the query was issued. In this paper, we aim to study gamers' search behavior, by using two-year query logs of the Persian search engine, Parsijoo (www.parsijoo.ir). To extract game searches, we relied on a list of popular video games and extracted queries containing their names. In order to complement this initial list, we also used the frequent words that are used for video game searches that were missing in our initial seed terms. We decided to set our research goals and answer the following research questions:

- RQ1: What is the temporal distribution of game searches and how do they differ from the general ones?
- RQ2: How do users specify the game version they are looking for?
- RQ3: What type of game categories are searched more often?
- RQ4: What platforms are mostly searched in game searches?
- RQ5: How do the game search sessions differ from the general ones?
- RQ6: What type of web pages are mostly clicked by users in game searches?

The overall purpose of this research is twofold. First, we aim to analyze how gamers use search engines. Second, we try to understand how can search engines may benefit from these answers in order to provide better services for their users. The remainder of this paper is organized as follows. Section 2 presents some of the related works. Section 3 describes the experimental setting and how we used Parsijoo query logs for our analysis. Section 4 discusses the results obtained. Finally, Section 5 outlines some conclusions and future work.

2 Related work

Related work to ours can be divided into two groups. The first group of researches, use search engine query logs to study users' behavior on general or specific topics to improve search engine effectiveness. The second group, is mostly related to study the effect of these video-games on human behavior. In this section, we try to cover some of the related researches to both of the categories.

Different works have been presented to understand users' search behavior. Query log analysis is a well-known research topic in the area of information retrieval and data mining which leverages the recorded information in the search engine query logs to study real users search behavior and based on the patterns found in their behavior enhance search engine effectiveness. One of the earliest works on query log analysis was proposed by Broder et al. [4]. In this work, the authors introduced a new taxonomy for web searches consisting of three categories: navigational, informational and transactional where for each category,

different retrieval results may be presented. Query logs were also used to understand the different type of queries, for which search engines failed to provide good results. For instance, Bendersky et al. [3] brought attention to the long queries and studied issues involved with using them effectively. They showed that user click behavior is correlated with the query length. In turn, Kravi et al. [11] studied multi-click queries where the same user may click more than one page in the same search session. They analyzed both queries and clicked pages, in addition to proposing a basic classifier to predict these types of queries.

While many previous researchers studied general web searches, there has been many attempts to explore specialized searches. For example, the study by Herskovic et al. [10] focused on PubMed query logs which is an interface to MEDLINE, the largest biomedical literature database. They studied search features such as length of query and session, size of the results sets and use of Boolean operators. The results obtained showed that users tend to issue a variety of queries on a broad range of topic most of which are informational. Another specialized research on query logs was conducted by Li et al. [12] where they studied academic searches focusing on the failures in these type of searches and providing an algorithmic solution for it. The last example to mention is the work of Spina et al. [18] where searches related to job and talent were investigated. In this work, the query log records of a vertical search engine, SEEK, were used to explore job searches observing different users' behavior compared to general searches. Despite all these works on the specialized searches, we believe no previous research so far has attempted to characterize game searches. By exploring web query logs of general search engine and extracting game-related queries, we aim to understand whether the behavior of a gamer is different from other common searches. This might be very useful for search engines seeking to provide better search results and dedicated services to their users.

As mentioned earlier, the second group of researches related to ours, are those regarding video games. These researches are mostly focused on how playing video games may affect human behavior. For instance, the recent experimental study by Hasan et al. [9] on 70 people showed that there is a stronger evidence of short-term violent video game effects than of long-term effects. In turn, Granic et al. [8] focused on the benefits of playing video games, mainly targeting motivational, emotional and social benefits of gaming. While most of these studies were conducted by user study with profiling users' searches, query logs can also be used on this matter. However, in this research we are not focusing on this issue and it can be considered as a potential future work.

3 The Video Game Search Dataset

To conduct our analysis, we resort to a Persian general-purpose search engine, named Parsijoo, in particular to a query log consisting of 27M queries and corresponding users' interaction collected during two consecutive years, from March 2015 to March 2017. Transaction log in web search engines is a record of the communication between users and search engines during a search session. The search log format is similar to the extended file format containing data such as user query, query issue time and clicked URLs. Table 1 shows these transaction logs with field SessionId, the submitted query, query issue time, Search type, Clicked URL, Click Time and Mobile. (note: queries are translated from Persian to English. Also due to space limit, URL and SessionID are shortened.) The "Mobile" column, shows if the query was issued from a mobile device (True value) or from a PC. Also, Parsijoo assign each Session Id when the browser is opened and a connection is made between user and the search engine. The same Session Id will remain for the user until the browser is not closed and the connection is not ended. As it may be possible that the user does not close the browser, we decided to partition the sessions into more realistic search session when there is no interaction between the user and the search engine for 60 minutes or more.

To extract game-related queries, we used the same keyword-based approach as Mansouri et al. [14]. In the mentioned research, editors were asked to manually construct a list of phrases for each of 100 events that were selected to study spiky events. Queries were then matched against these lists. In our research, we used a list of 600 popular games (published by a national computer and video games foundation), from online games, such as "Clash of Clans", PC games such as "Call of Duty", PlayStation games like "Mortal Kombat" to old video games like "Super Mario". Each popular game was then issued into the query log records. Any query containing all of the terms of the popular games is considered as a game search related query. For

Table 1. Parsijoo transaction logs for two search sessions with game queries.

Session ID	Query	Issue Time	Clicked URL (Site shown)	Click Time	Mobile
B12A561CB6	Clash of clans update	2016-12-22 21:45:30			True
B12A561CB6	Clash of clans Christmas update	2016-12-22 21:46:02			True
B12A561CB6	Clash of clans Christmas update iOS	2016-12-22 21:47:10	cafeapple.net	2016-12-22 21:47:23	True
C31B4235A7	mortal kombat instructions	2016-03-03 23:15:34			False
C31B4235A7	mortal kombat finishing moves	2016-03-03 23:16:02	zoomg.ir	2016-03-03 23:16:38	False
C31B4235A7	mortal kombat Scorpion moves	2016-03-03 23:20:39			False
C31B4235A7	scorpion mortal kombat finishing moves ps4	2016-03-13 23:21:00	gamefa.com	2016-03-13 23:21:24	False

instance, the query "release date of call of duty" is considered as a game-related query as it contains "call of duty".

In the next step, we explored the extracted records to identify other frequent terms that are used in game search queries. For this purpose, we start by considering the top-100 frequent terms but selected only those related to the game search task. Eventually, 50 more terms such "online games", "Lan game" or "pc multiplayer game" to name but a few, were selected, thus totalizing 650 seed terms. With this method, we ended up selecting a considerable number of 372,508 game search records from the query logs, corresponding to 253,516 search sessions which we would regard them as game search sessions. A game search session does not necessarily start with a game-related query, but contains at least one query related to video games. To compare the game searches with general searches (those not related to video game), with the same constraint about search session length, we randomly selected search sessions until we reach the same amount of records as game searches (372.5K).

4 Experimental Results

In this section we present the results of our analysis on the game-related queries. All the concepts that we decided to study were determined after our initial exploration on game queries and the patterns observed. First the top-2000 frequent queries were selected to find patterns. As temporal expressions and name of platforms were used in nearly 62% of these queries, we decided to explore how users used them in their video-game queries. The same approach was used for the clicked-pages. By looking at the URL (not the content) of top-1000 clicked pages, our assumption was that users mostly click on game-oriented and application download websites rather than other web pages. Finally, we selected a set of 100 random game search sessions to observe the number of queries and based on our previous knowledge on general searches, the number of queries was nearly the same, however in the query reformulation, some patterns could be observed. The following subsections will verify if our initial assumption about video game-related queries were correct or not.

4.1 Temporal distribution

Understanding changes in queries over time can help search engines providing better web search services. For instance, Shoukouhi and Radinsky [17] proposed a time-sensitive approach for query auto-completion, ranking the candidates queries according to their forecasted frequencies. Another benefit of knowing the temporal distribution of special topic queries is advertisement. Many game producers rely on introducing their games to consumer by advertising them online. By knowing when the specific audience in the market is available (doing the web search), search engines can have better timing for their ads. In sponsored search,

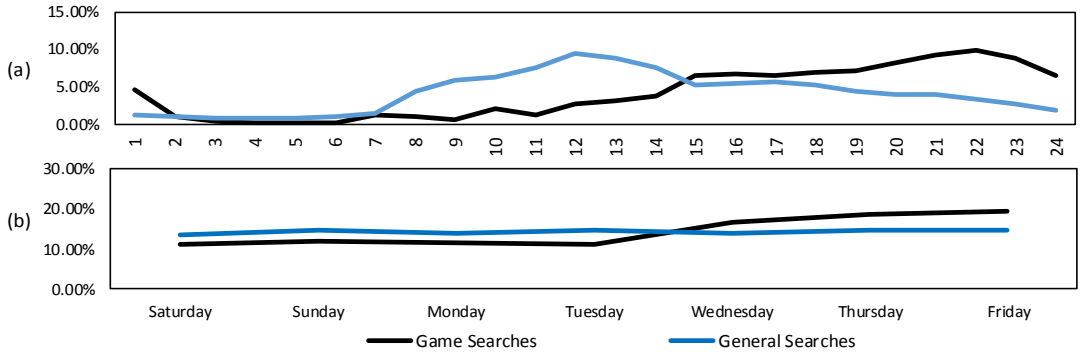


Figure 1. The distribution of game search and general web search queries during day hours (a) and week days (b).

an advertiser only pays if its ads get clicked by the user [1], and putting video game ads at these time will increase the chance of having them clicked.

In the research of Beitzel et al. [2], authors studied how the popularity of topically categorized queries changes on a daily, weekly and monthly basis. Their study on game category showed that, the number of related queries were, more or less the same, regardless the hours of the day and the day of the week. However, based on our observation, these queries might be more issued at the end of the day and weekends when talking about game search queries. To conduct this analysis, we used the query issue time and plotted the frequency of game and general searches over the hours of the day and the days of the week. The result of our experiment, on temporal distribution of game queries when compared to general ones, is shown in Figure 1 by daily hours (a) and week days (b).

Exploring the distribution of game queries, these queries are mostly issued during the night when people are usually back at home and willing to spend their leisure time. The frequency of game searches, starts to increase from 15h and reaches its peak at 22h when approximately 25% of the game queries is issued. In contrast, the frequency of general searches increases from the morning to the noon and then gradually decreases as it approaches the end of the day.

Our further analysis on the distribution of game and general searches during week days indicates that, while the frequency of general searches, remained more or less the same during the week day, games searches were more issued during the weekend (in Persian calendar, Thursday and Friday are the weekend days, and Saturday marks the beginning of the week). From Saturday to Tuesday, both game and general searches show the same constant trend. But while general searches continue their stable frequency, game queries show a sudden increase in the number of game queries posed by users from Wednesday (which anticipates the beginning of the weekend) to Friday when it reaches its peak with approximately 20% of the game search queries issued.

We also studied temporal distribution for some of the popular games. For many of the games the time-series built on their query frequency had multiple spikes. Queries with multi spikes in their frequency time series are usually temporal ambiguous. Temporal ambiguous queries have been studied by several researches [5, 15, 16], indicating the need for different retrieval methods. For games such as PES (pro evolution soccer), there were periodic spikes around release dates, with multiple queries regarding the release date and new features. On the other hand, for games such as “Clash of Clans”, there were aperiodic spikes in the frequency time series. These spikes were due to updates on the game and users were looking for update links. For both type of periodic and aperiodic spikes, search engines should try to adopt different retrieval method based on query issue time [14]. Figure 2, shows the time series built on the monthly frequency of queries related to games “PES” (a) and “Clash of Clans” (b).

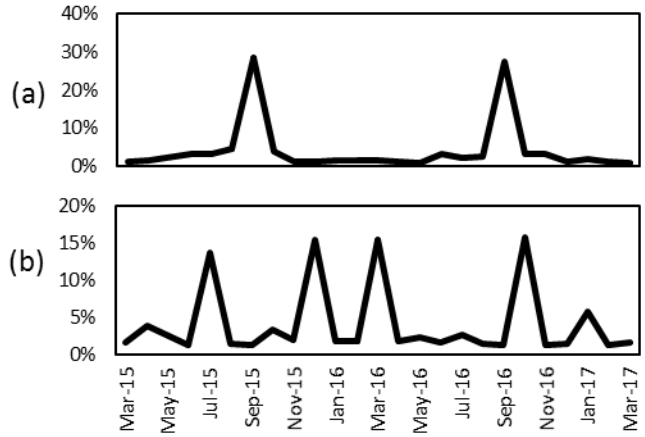


Figure 2. Distribution of queries for games PES (a) and Clash of Clans (b) in Parsijoo query logs.

4.2 Game Version Specification

Users mostly try to disambiguate their query by adding extra terms to their initial queries. As the previous section showed queries related to games, may have temporal ambiguity; that is for the search engine, it is not clear which of the time ranges are of the interest for the user. One way to solve this problem is to add temporal expression to the query, specifying the desired time range. Therefore, we studied if users have used a temporal expression when specifying their game-related query. To study this matter, we used a Persian Time tagger, ParsTime [13].

Our analysis shows that in 19.32% of game queries users included temporal expressions in their queries of which (based on the query issue time), 88.3% of the users were looking for the most recent version of the game, while only 9.1% were looking for older versions and a tiny portion of 2.6% were interested in future versions. This emphasizes the need for search engines to adopt special retrieval techniques that take into account the freshness of the results when talking about game search queries, which is a challenging task [6]. Another point is that, the version of games can also be specified by numbers, specially the roman number such as (I, II and III). Considering the numbers as the version of the games, 25% of the users have specified the version of the game they were looking for. However, these numbers cannot be studied to check the user target time.

4.3 Game Categories

There are different categories of video games and users may share a same taste. For instance, a type of video game can be popular in a specific region or be popular between people of a certain age range. Also a user might be only interested in a certain type of video games. For game search engines with game recommendation module, the validity of this assumption is crucial as they can benefit from one such information in a task such of query suggestion. In our query logs, user profiles are not known; that is, we do not have the history of a user search. However, we try to study if users attempt to change the game category within a same search session. Also by having user IP, we can study the distribution of video game categories in different regions. To do so, we begin by asking 3 editors of the search engine crowdsourcing team, to specify the genre of 500 selected video games (in our seed from Section 3) based on their own knowledge or

Table 2. The distribution of different game categories searched by users.

Category	Distribution(%)	Category	Distribution(%)	Category	Distribution(%)
Sports	26.30	Racing	9.90	Puzzle & Word	3.90
Action	13.40	Adventure	8.50	Other	7.40
Strategy	25.70	Board & Card	4.90		

on the game description, according to google play game categories. An inter-rater reliability analysis using the Fleiss Kappa statistics [7] was performed to determine consistency among the editors. Overall, the annotators obtained about 0.78 of agreement level, which represents a high agreement between editors. Overall, in 81.6% of game queries the game was specified (which can be due to our method for extracting game queries), The distribution of different game categories is shown in Table 2.

As this table suggests certain categories such as sports, strategy and actions games are more popular than the other categories. Our analysis on game search sessions, shows that in 14.5% of game sessions, users tend to search for different video games. In these search sessions, however, only in a minority of 0.9% sessions, users changed the category of the game there are looking for. This indicates that in game search sessions, users mostly target certain category of games and if search engines are planning to consider query recommendation module, games of the same category should be suggested.

The next experiment was conducted on the popularity of game categories within cities. The result of our analysis shows categories such as actions and strategy games were dominantly issued in industrial cities (84.8% on average). On the other hand, categories such as "Puzzle & Word" and "Board & Card" games were not popular in these cities (16.2% on average). One such information can also be used by game oriented web pages to suggest games based of user's location.

4.4 Game Platforms

If users do not specify their desired game console, search engines cannot distinguish which console the user was targeting thus impeding a diversity of the results based on the platforms. In this section, we study how often users specify the device for which they are searching games. As there has been noticeable increase on number of mobile video games, we also plan to study if there is a correlation between the device from which the query was issued (mobile or PC) and the platform users are looking for. Overall, in 47.4% of the game queries, users have specified the platform. Table 3 shows the distribution of the devices specified by users in games queries (over all the queries with device specified). The other category refers to platforms such as Sega and Atari. Also it should be mentioned that the mobile category includes queries with "Android" and "IOS". Games for Personal computer and Mobile were mentioned in more than half of the game queries in which the platform was specified.

By using "Mobile" field from query logs, it was observed that 92.8% of the game queries with mobile as the specified platform, were issued from mobile and 84.7% of the queries related to PC platform were issued

Table 3. The distribution of game platforms specified by users in game queries.

Platform	Distribution(%)	Platform	Distribution(%)
PC	27.4	PlayStation	14.9
Online	24.3	Mobile	26.7
Xbox	3.3	Other	3.4

from desktop PCs. This clearly state that, the device from which the query is issued, can be used for query suggestion; that is to say, if the user is posing a query related to game by a mobile device (Android or other operating systems), the results should be related to that operating system. Likewise, if the search is issued on a personal computer, then video games related to PC should be suggested.

4.5 Search Sessions

So far we have mostly focused on the game queries. Instead, in this section we study the game search sessions and the queries issued in these sessions to better understand game search sessions. First, we study how user search effort is different in game searches compared to general ones. For this, we considered average query length and average query per search session. The comparison between query length of game and general web queries shows that game searches are about 2.9 terms longer than general web searches, with an average of 5.4 terms per query (while only 2.5 terms on general queries). This might be explained by the fact that users may specify the name of the game, game console, and game's version and also use terms such as "download" in their queries. To better understand the difference between query length in game and general searches, we drew the boxplot of distribution of the queries for both types of searches in Figure 3. By looking at this figure we can observe that most frequent game search queries have a length between 4 to 6 terms, with queries with more than 9 terms being very rare.

Next, we study the average number of queries posed by a user in each search session in order to understand whether there is any difference between game and general web searches. Our analysis on the average number of query per session in game and general searches shows that users pose fewer queries in game search sessions compared to general search sessions. The average query per session in game searches is 1.46 queries per session which though quite similar is less than the 1.68 queries posed on general searches. This indicates that users may put less effort in game searches as they can find the appropriate result for their information need more quickly. Overall, our analysis revealed that at the current moment, users put less effort in game searches compared to general ones which may indicate that search engines can provide appropriate result for game searches. The query reformulation mostly happened 24% of game search sessions. In these search sessions adding words reformulation (where user add one or more new term to the previous term) dominantly occurred for 78% of times.

Going into the details, in search sessions with adding word reformulation 67.3% of queries were reformulated by adding the platform names where 22.9% of reformulation were done by adding the temporal expressions and numbers to the initial queries. For instance, in a search session the user initial query was "Clash Royal update", reformulated by adding temporal expressions to "Clash Royal December Update" and then reformulated by adding platform to "Clash Royal December Update for Android".

Another example is the search session in which user first issues the query "Hay Day download" and without clicking any results issue another query by adding the platform "Hay Day download IPhone" and by clicking

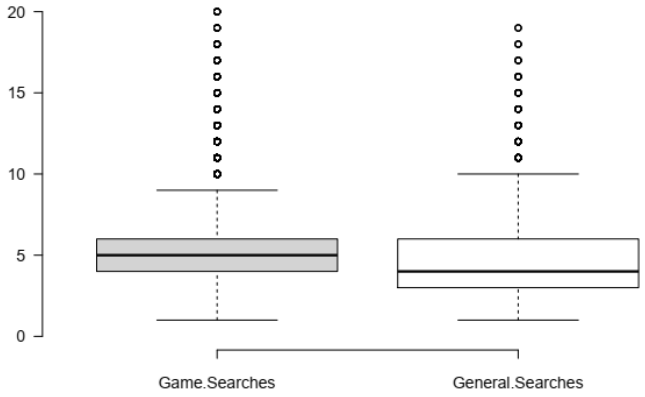


Figure 3. Boxplot of query length distribution for game and general web searches.

on two pages finished the search session. This shows that users may not specify the platform in their initial query, hoping they can find their desired results in the first try.

4.6 Clicked Web Pages

Our final analysis was on the type of clicked pages by users in game searches. Our initial observation on the clicked pages showed that certain web pages related to video games and pages providing download links for them are more clicked. Among the top-500 clicked web pages which account for 82% of total clicked pages in video game searches, 55.3% of web pages were game-oriented, 24.6% were application download web sites, 15.8% were websites dedicated to mobile application downloads and 4.3% of the web pages belong to other categories (this was done by only one annotator based on the simplicity of the task.) A deeper analysis enables to conclude that within the top-10 clicked web pages, 6 out of 10 are game-oriented, 3 refer to mobile applications and one to download web pages. This information may be used by search engines to give higher ranks to web pages belonging to these three categories for game related queries.

5 Conclusion

In this paper the web queries related to video games were explored. The analysis on queries, search sessions and clicked pages showed that search engines should provide different approaches for both retrieval and search services they provide for game searches. We studied different aspects of game searches and provided solutions for search engines to improve their effectiveness. For instance, the temporal distribution of video game searches is different from general searches usually being done at the weekends and at the end of the day. One such information can be used by search engines who plan to advertise video game, indicating appropriate time to put on the ads to meet the expected audience. Another example of our findings is the correlation between the platform users are targeting and the device the query was issued from. 92.8% of the game queries with mobile as the specified platform, were issued from mobile and 84.7% of the queries related to PC platform were issued from desktop PCs. This information can be used by search engines to retrieve the related pages based on the device from which the query was issued; that is if the game query is issued from mobile device, perhaps the user is looking for the mobile version of the game. As part of our future work, we plan to explore video game searches deeper and also develop a vertical search engine for video games.

ACKNOWLEDGMENTS

This work is funded by the ERDF through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT as part of project UID/EEA/50014/2013.

REFERENCES

[1] Bai, X., Ordentlich, E., Zhang, Y., Feng, A., Ratnaparkhi, A., Somvanshi, R., & Tjahjadi, A. (2018). Scalable Query N-Gram Embedding for Improving Matching and Relevance in Sponsored Search. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 52-61). ACM.

[2] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized web query log. Journal of the American Society for Information Science and Technology, 58(2), 166-178.

[3] Bendersky, M., & Croft, W. B. (2009). Analysis of long queries in a large scale search log. In Proceedings of the 2009 workshop on Web Search Click Data (pp. 8-14). ACM.

[4] Broder, A. Z., Carmel, D., Herscovici, M., Soffer, A., & Zien, J. (2003). Efficient query evaluation using a two-level retrieval process. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 426-434). ACM.

[5] Campos, R. N. T. (2013). Disambiguating implicit temporal queries for temporal information retrieval applications.

[6] Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., ... & Diaz, F. (2010). Towards recency ranking in web search. In Proceedings of the third ACM international conference on Web search and data mining (pp. 11-20). ACM.

[7] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.
[8] Granic, I., Lobel, A., & Engels, R. C. (2014). The benefits of playing video games. American psychologist, 69(1), 66.
[9] Hasan, Y., Bègue, L., Scharkow, M., & Bushman, B. J. (2018). "The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior": Corrigendum.
[10] Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: analysis of a typical day's query log. Journal of the American Medical Informatics Association, 14(2), 212-220.
[11] Kravi, E., Guy, I., Mejer, A., Carmel, D., Maarek, Y., Pelleg, D., & Tsur, G. (2016). One query, many clicks: Analysis of queries with multiple clicks by the same user. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 1423-1432). ACM.
[12] Li, X., Schijvenaars, B. J., & de Rijke, M. (2017). Investigating queries and search failures in academic search. Information Processing & Management, 53(3), 666-683.
[13] Mansouri, B., Zahedi, M. S., Campos, R., Farhoodi, M., & Rahgozar, M. (2018). ParsTime: Rule-Based Extraction and Normalization of Persian Temporal Expressions. In European Conference on Information Retrieval (pp. 715-721). Springer, Cham.
[14] Mansouri, B., Zahedi, M. S., Campos, R., Farhoodi, M., & Rahgozar, M. (2018). Understanding User's Search Behavior towards Spiky Events. In Companion of the The Web Conference 2018 on The Web Conference 2018 (pp. 1763-1769). International World Wide Web Conferences Steering Committee.
[15] Mansouri, B., Zahedi, M. S., Rahgozar, M., Oroumchian, F., & Campos, R. (2017). Learning temporal ambiguity in web search queries. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2191-2194). ACM.
[16] Shokouhi, M. (2011). Detecting seasonal queries by time-series analysis. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 1171-1172). ACM.
[17] Shokouhi, M., & Radinsky, K. (2012). Time-sensitive query auto-completion. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 601-610). ACM.
[18] Spina, D., Maistro, M., Ren, Y., Sadeghi, S., Wong, W., Baldwin, T., ... & Zobel, J. (2017). Understanding user behavior in job and talent search: an initial investigation. In SIGIR Workshop on eCommerce (pp. 1-5). SIGIR Workshop on eCommerce.

Wiki Workshop Chairs' Welcome

It is our great pleasure to welcome you to the *Wiki Workshop at the Web Conference 2019*. The goal of this workshop is to bring together researchers exploring all aspects of Wikimedia projects such as Wikipedia, Wikidata, and Wikimedia Commons. With members of the Wikimedia Foundation's Research team on the organizing committee and with the experience of successful workshops in 2015, 2016, 2017, and 2018 we aim to continue facilitating a direct pathway for exchanging ideas between the organization that coordinates Wikimedia projects and the researchers interested in studying them.

We received 22 paper submissions from all around the world covering a broad range of topics. Our program committee evaluated them regarding relevance, quality, and novelty, selecting 13 to be presented as posters and/or oral and published in the workshop proceedings, and numerous others for poster presentation only.

In addition to these contributed posters, we are looking forward to hosting 5 high-profile invited speakers (in alphabetical order):

- **Timnit Gebru,** Research Scientist at Google's AI Ethics Lab
- **Erica Kochi,** Co-Founder of the Unicef Innovation Unit
- **Jure Leskovec,** Associate Professor of Computer Science at Stanford University.
- **Neil Thompson,** Research Scientist at MIT CSAIL, Visiting Professor at Harvard.
- **Denny Vrandečić,** Ontologist at Google Knowledge Graph.

We hope that you will join us for these speakers' presentations and for the lively discussions that are certain to ensue around the most pressing issues and the most promising opportunities faced by Wikipedia, Wikidata, and other Wikimedia projects.

Miriam Redi
Workshop Co-chair
Wikimedia Foundation

Robert West
Workshop Co-chair
EPFL

Dario Taraborelli
Workshop Co-chair
Wikimedia Foundation

Wiki Workshop Organization

Workshop/Track Chairs:
Miriam Redi (Wikimedia Foundation)
Robert West (EPFL, Switzerland)
Dario Taraborelli (Wikimedia Foundation)

Program Committee:
Lucas Dixon (Jigsaw)
Besnik Fetahu (L3S Hannover)
Andrea Forte (Drexel University)
Gary Hsieh (University of Washington)
Yiqing Hua (Cornell University)
Isaac Johnson (Wikimedia Foundation)
Os Keyes (University of Washington)
Markus Kroetzsch (University of Dresden)
Florian Lemmerich (RWTH Aachen University)
Lauren Maggio (Uniformed Services University)
Jonathan Morgan (Wikimedia Foundation)
Daniela Paolotti (ISI Foundation)
Tiziano Piccardi (EPFL)
Dario Rossi (Huawei)
Diego Saez-Trumper (Wikimedia Foundation)
Markus Strohmaier (RWTH Aachen University)
Nithum Thain (Jigsaw)
Michele Tizzoni (ISI Foundation)
Morten Warncke-Wang (Wikimedia Foundation)
Joe Wass (Crossref)
Ramtin Yazdanian (EPFL)
Amy Zhang (MIT)

Open Personalized Navigation on the Sandbox of Wiki Pages

Chuankai An*
chuankai@cs.dartmouth.edu
Department of Computer Science, Dartmouth College
Hanover, NH, 03755

Daniel N. Rockmore†
Department of Computer Science and Mathematics,
Dartmouth College
Hanover, NH, 03755

ABSTRACT

In this paper we present a proof-of-concept of a visual navigation tool for a personalized "sandbox" of Wiki pages. The navigation tool considers multiple groups of algorithmic parameters and adapts to user activity via graphical user interfaces. The output is a 2D map of a subset of Wikipedia pages network which provides a different and broader visual representation – a map – in the neighborhood (according to some metric) of the pages around the page currently displayed in a browser. The representation schema includes the incorporation of a kind of transparency in the algorithmic parameters affecting the presentation of the landscape visualization, which in turn enables the delivery of a personalized canvas, designed by the user. A case study shows the combination of four different sourcing (i.e., identification and extraction of the neighboring pages) rules and three layouts over the same Wikipedia subnetwork. The basic schema is readily adapted to other search experiences and contexts.

CCS CONCEPTS

• **Information systems** → **Wikis**; • **Human-centered computing** → *Graph drawings*; • **Computing methodologies** → Network science.

KEYWORDS

Wikipedia; Network Visualization; Transparent Navigation

ACM Reference Format:
Chuankai An and Daniel N. Rockmore. 2019. Open Personalized Navigation on the Sandbox of Wiki Pages. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316755

1 INTRODUCTION

Wikipedia is an important source of information [30]. For many people, going to Wikipedia is just the first step in an information search task. A standard search trajectory would then take place, realized as a sequence of clicks, effectively something of a constrained, yet still "random" walk from the Wikipedia page starting point, alike at least in spirit to the "random surfer" model that gave birth to PageRank [9] (and Google) whose depth and penetration of the space of relevant webpage resources can and does depend on many contingent characteristics. Regardless of the starting point, in this click-by-click revealing of the relevant (one hopes!) knowledge, it may be easy to miss or get distracted away from the original motivation for inquiry. More broadly, in such a blinded navigation the user is unaware of the way in which the webpages she visits relate to one another. Inadvertently she may be stuck in cul-de-sac of narrowly defined information or strayed very far from her initial search goal. It is with this in mind, that we take on and suggest an alternate option, one that promotes a notion of visual search, that presents a map-like visual summary of the page neighborhood, thereby possibly promoting a broader field of vision in the search process and highlighting different criteria for navigation.

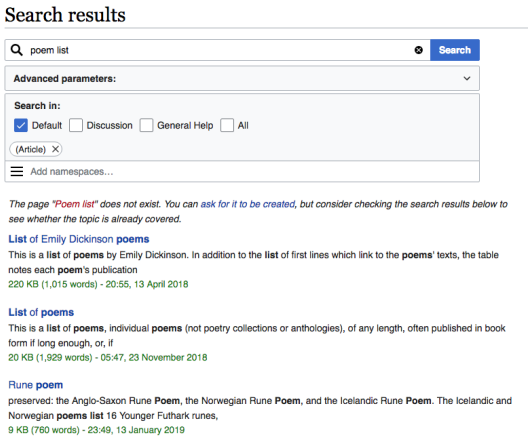
We propose a visual navigation tool for Wikipedia based network visualization, which allows users to select their preferred query target as the root page, and visualizes the local "sandbox" of related pages in the form of a 2D map on a "canvas" (viewing platform). Our motivation arises from the user experience of standard query on Wikipedia. Figure 1a shows a list of related pages when the query input is not exactly matched to a Wikipedia page. More often, Wikipedia will load a new page in the browser or redirect it to a similar one as Figure 1b. We believe a visual navigation tool might be a useful broadening of our verified knowledge boundary during browsing better than such a list of results or unexpected redirection.

For example, a series automatically updated maps of the surrounding pages in the network space could display a broader view of the information space that both illustrate the distance among those nodes (pages) while also providing some sense of context for the material on the page. A visual navigation tool based on Wiki page networks could also facilitate a user's understanding of the local network structure, and would bring more transparency to the query results. While the network structure articulates the link relationships between pages, the use of other kinds of metadata (from the user and other users as well from the webpages) raises the possibility of creating a non-link distance structure (metric) for the neighborhood, and with that, new possibilities for display and user interaction. User response to the 2D Wiki map might also offer interaction data for user-behavior oriented research projects. While the focus of this article is on the Wikipedia environment, the general framework of user-controlled network navigator is not limited to Wikipedia corpus. For example, the dynamic graph visualization may also work as a recommender for online shopping or the World Wide Web as a whole.

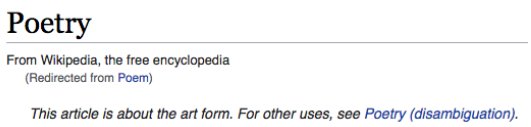
In this paper, we present some initial ideas around the design of a personalized visual navigation system on Wikipedia. Generally, the data flow starts from a seed Wikipedia page. A "sandbox" of related pages is defined by a distance threshold on the Wikipedia

*Corresponding Author
†External faculty of the Santa Fe Institute, Santa Fe, NM, 87501

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316755



(a) Query result: a list of related Wikipedia pages.



(b) Redirection to a new page.

Figure 1: Interfaces of query results on Wikipedia.

page network, in which a directed edge from page A to page B means page A cites B in the HTML source file. The navigator will interact with users to get the desired algorithmic parameters of the visualization to be personalized. Behind the user interface, diverse algorithms implement the tasks of node filtering, coordinate computation and edge selection due to a limited size of screen. Though researchers have intended to diversify the user experience of Wiki with visual effects [13, 24, 27], our contributions include the possibility of real-time updated visual navigation that responds to users browsing in the open space of Wikipedia pages, and the fact that nodes on the screen are determined by the personalized algorithmic parameters directly set by users. As users are often only aware of the pages that comprise their browsing paths only relatively "blind" to any "surrounding" ones, we hope our design of such an immersive visual navigation would make for a more useful Wikipedia search experience.

Section 2 surveys related work about Wikipedia tools and network navigation. In Section 3, we propose a data pipeline of the visual navigation system and the options of three groups of algorithmic parameters. Section 4 provides the details of implementation. Section 5 illustrates the final output of network visualization of a query word, as well as the comparative analysis of the effects of algorithmic parameters. Section 6 briefly describes the design of user study. Finally, Section 7 concludes the initial findings and directions of further development.

2 RELATED WORKS

Several projects have focused on Wikipedia navigation, as relates to efficient browsing. Lamprecht [19, 20] discussed the influences of Wikipedia navigation policies and the structure of Wikipedia pages network. Odor [24] presented the evolution of Wiki hyperlink networks to aid navigation and understanding. Another automatic tool [27] aimed to generate info-boxes for Wikipedia pages from a Wikipedia knowledge graph. In a different direction, Leibon et al. [22] show how the Wikipedia pages around a given topic – e.g., mathematics – can support a metric and thus structure of a hyperbolic geometry and with that, enables the construction of geodesics (paths in the Wikipedia space) that optimally guide a user's viewing experience (the use case of the paper is the MathWikipedia). Figshare [1] provided content-based embeddings learned from Wiki corpus as the navigation vectors on a 2D plane. Cartograph [29] enables the presentation of a vast map of Wikipedia pages with the embeddings learned from neural networks. The last of these differs from our proposed navigation schema which when fully realized would allow users to change the visualization and any underlying metric supporting the visualization.

Also related is work on semantic annotation and some applications of collaborative filtering applied to Wikipedia data. IkeWiki [28] and SweetWiki [10] made the inherent structure of a Wikipedia page accessible to users and computing machines via annotations derived from semantic methods (e.g., RDF and conceptual graphs). A visual analytics framework [13] illustrated how editors could work together for a public visualization of Wikipedia data.

Researchers also have been working on diverse kinds of Wiki tools to improve knowledge transfer and user experience. Harder et al. [17] designed a new measure to model and display the degree of "verifiability" of a Wikipedia page and implemented a demo in a Chrome browser extension. A visual article development tool [15] explored editor interaction history to deal with disagreement. Balaraman [7] proposed a new metric to describe the relative completeness of Wikipedia data. Gundala [16] reported the initial progress about predicting hyperlinks between pairs of non-connected pages that are helpful for search navigation. WikiTrails [26] provides a tracking system of visited Wikipedia pages to facilitate the understanding of Wikipedia content structure. Omnipedia [8] visualized multi-language editions of the same Wikipedia page via colorful circles in different sizes based on an article alignment algorithm, but it ignored network analysis. SuggestBot [11] proposed a link recommendation framework to match people with suitable editing tasks on Wikipedia.

Lastly, there is now a growing body of experimentation with the digital interfaces for searching and exploring traditional information materials, specifically for the interaction with libraries. An interesting example of this is the Harvard Stacklife project[1], that aims to bring back to online library search the missing – and bemoaned – loss of the serendipity of browsing the stacks that occurs when going to retrieve a book of interest. It is in the spirit of such exploratory serendipity that we present the work in this paper. Another interactive graph [21] allows users to set up the weights of link structure measures and textual similarity for a 2D map of legal documents, from which researchers explore how new opinions influence the search behavior of judges and litigants and thus affect the law.

[1] http://stacklife.harvard.edu

3 DATA PIPELINE

In this section, we present the data pipeline starting from a seed Wikipedia page to the visualization of related pages. The following steps combine user interface and algorithm-based computation for a personalized and transparent visualization. In this case, transparency means that users would know how the thing they are looking at is made, while customizability means that users have the power to directly change the input parameters in our proposed navigation system. We hope the customizability and the idea of network space visualization could make the navigation system more transparent.

Wikipedia Seed Selection. The "seed page" represents the user-defined center or starting point for a neighborhood of Wikipedia pages of interest for a given topic.

Wikipedia pages crawler. If a Wikipedia page cites another one in the main context, they are a pair of linked nodes in the network of Wikipedia pages. Among the billions of Wikipedia pages, in this preliminary proof-of-concept study we work with only a very small subset and limit the range of the crawler with some threshold on the distance between a candidate page to the seed.

Parse sandbox structure. Given a seed page, the downloaded subset of Wikipedia "nearby" pages is our "sandbox". Our processing and analysis do not edit their content. This step aims to build the network of the observed local Wikipedia pages around the seed. There are various options for thresholding the neighborhood (e.g., all pages within some fixed linked distance of the seed). A Wikipedia (sub)-network not only contains the surrounding nodes, but also the edges among them. Here we define the weight of an edge between two nodes as the number Wikipedia pages that cite both of them. For a given node, this weight enables a sorting of its direct neighboring nodes (i.e., with a distance of one) in the network.

Set algorithmic parameters. Different from traditional digital-art based user interface (UI) design, here we propose a framework for algorithmic visualization for a (sub-) network of Wikipedia pages. It contains three groups of parameters, set by users, to make the visualization more transparent:

(1) The rule of nodes sourcing and ranking. Here we apply four different methods: semantic content-based similarity, graph structure, collaborative filtering of users browsing and an overall *PageDist* (cf., [21]) metric derived from link and content similarity. The navigation tool could display a limited number of nodes within a canvas, compute the internode distance matrix and then use that for node placement/visualization. For example, we compute the distance between two Wikipedia page titles according to the word vector representation GloVe [18]. It is also possible to sort the neighboring nodes with some network science features (e.g., node degree, PageRank centrality). To simulate collaborative filtering, we assume that the frequency of concurrence on a third page is proportional to the probability of users preference for the two pages. The PageDist [21] metric considers the commute distance in a transition matrix, which is derived from both in-out link structures and semantic similarity of texts. All the sourcing methods are independent with the link structure of the downloaded pages in the sandbox. In the previous step of crawling, we have applied a distance threshold. If the navigator serves the whole Wikipedia network without a radius in the crawling step, those sourcing methods could narrow down the range of candidates as well.
(2) Definition of "nearby" nodes on the canvas. According to the selected sourcing and ranking method for a small number of candidate nodes to display with the seed on a canvas, we could sort all surrounding nodes according to their feature similarity to the root page. Therefore, for any two nodes appearing on the canvas, their relative proximity to the center node would be in accordance with their rankings in the sorted list of their feature similarity to the root page. Another definition of proximity comes from the result of a user preference predictive model, where the neighboring nodes with a larger probability of preference will be closer to the center node. Those two settings may be in accordance with each other, but sometimes a user might explore some new pages rather than the most similar one. This might be especially true when looking for information about particularly divisive or "charged" subjects.
(3) Layout of nodes. We try to locate the current Wikipedia page at the center of a canvas, except in the case of using the 2D multidimensional scaling [12] (MDS). If the second setting (i.e., closeness to the center) is defined by the feature similarity from the sourcing rule, the surrounding nodes should follow the order of their distances to the root page in the feature vector space. We implement spiral and spectral layouts to adapt to a ranking of the selected nodes. Both layouts point out the "close" neighbors and grant users the access to adjust parameters for their desired neighboring nodes. The assigned node coordinates in MDS match the idea of preserving between-node distances rather than the arbitrary design of spiral/spectral layouts.

Visualization. As a part of computational visualization, several factors might limit the actual effects, such as the size of the available screen (i.e., "canvas") to present the Wikipedia network, the number of pixels in a fixed size canvas (i.e., resolution), and the suitable number of nodes/edges. In addition, the location of a node should follow the general direction defined by the layout. Therefore, except for the MDS option, we first compute coordinates in a polar system, then transform it to the 2D plane coordinates. This step needs the help of an external visualization package which places nodes on a 2D plane at the accurate coordinates, so that users could present a non-standard yet desired layout on a canvas. Though many algorithmic terms are introduced in this tool, non-expert users could compare the differences in visualization and – with a little experience and/or training – adjust parameters for their preferred result.

Update the Sandbox. We hope the navigation system could incorporate user activities, such as hyperlink clicks, revisiting a page or long-time browsing. Once monitoring the above activity, the system should return to the second step to crawl some new Wikipedia pages, and update the Wikipedia network with the following steps, such as a new seed page and new selected neighbors. In this way, the navigation tool could extend to an open Wikipedia space and

gradually collect user preference records for other personalized services on Wikipedia.

4 SYSTEM IMPLEMENTATION

In this section we briefly introduce the implementation of a proof-of-concept visual navigation system for Wikipedia pages which follows the data pipeline defined in Section 3. It is developed in Python because of multiple efficient existing programming packages. Selenium [23] enables the detection of the current URL in a browser. Tkinter [4] offers the UI modules (e.g., input frames and radio buttons) for Wikipedia seed confirmation in Figure 2 and algorithmic parameters settings in Figure 3. With the input of a seed Wikipedia page, Urllib [5] downloads all the cited Wikipedia hyperlinks in the seed page with the help of a regression expression matching function. BeautifulSoup [2] facilitates the analysis of hyperlinks in local HTML files so that we could build the network of Wikipedia pages in the "sandbox". NetworkX [3] could place a node at the given coordinates in a 2D plane, so the navigator displays the same layout as what users choose (see Figure 3). Here we show examples of the visualization.

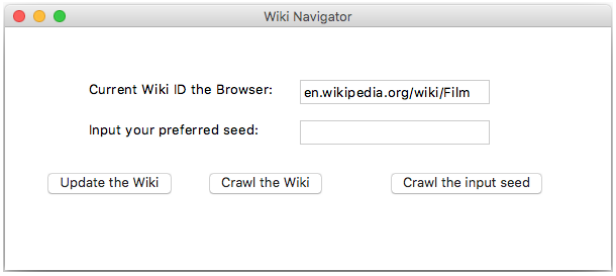


Figure 2: The user interface of Wikipedia seed selection. Users could input a seed or select the current one in a browser.

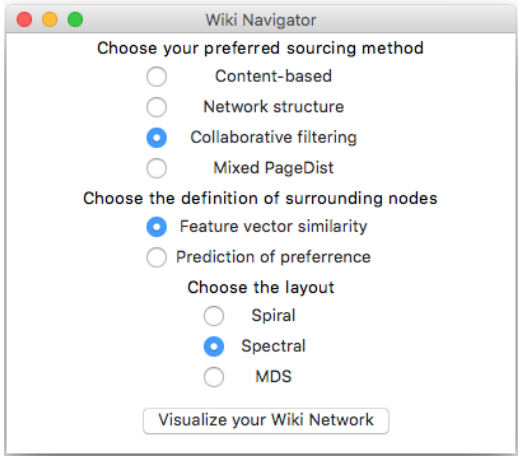


Figure 3: The user interface of algorithmic parameter selection for network visualization.

The implementation we describe above thus assumes an offline deployment to determine a subset of edges and links on the screen.

1	Film	36	Classical_Hollywood_cinema
2	Screenplay	37	Cult_of_personality
3	Documentary_film	38	Public_relations
4	Television	39	Principal_photography
5	Film_production	40	Color_motion_picture_film
6	Film_genre	41	Spectacle_(critical_theory)
7	Short_film	42	Script_breakdown
8	Art_film	43	Videography
9	Movie_studio	44	Main_Page
10	Independent_film	45	Film_industry
11	Sound_film	46	Cinematography
12	Silent_film	47	Special_effect
13	Soundtrack	48	Internet
14	Science_fiction_film	49	Visual_effects
15	Film_history	50	Post-production
16	Film_director	51	Storyboard
17	Film_editor	52	Film_score
18	Feature_film	53	Film_crew
19	Animation	54	Sound_effect
20	Film_release	55	Guerrilla_filmmaking
21	Film_editing	56	Filmmaking
22	Pitch_(filmmaking)	57	Streaming_media
23	Digital_object_identifier	58	American_Dream
24	Concentration_of_media_ownership	59	Film_treatment
25	News_broadcasting	60	Media_event
26	Shooting_schedule	61	Docufiction
27	Occupation_(protest)	62	Culture_industry
28	Cinema_of_the_United_States	63	Managing_the_news
29	Crowd_manipulation	64	Strike_action
30	Recuperation_(politics)	65	United_States
31	International_Standard_Book_Number	66	Daily_progress_report
32	Daily_production_report	67	Mainstream_media
33	Breaking_down_the_script	68	Screenwriting
34	Demonstration_(protest)	69	Political_satire
35	Roadshow_theatrical_release	70	Bollywood

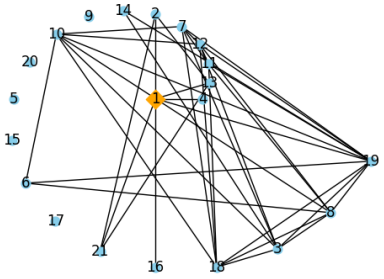
Table 1: Dictionary of the nodes selected by four sourcing methods.

Extensions of this simple approach may include (1) a much wider range of online Wikipedia pages around the seed page (2) a combination of more advanced algorithmic settings without too much time cost.

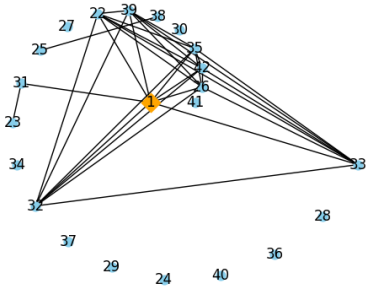
Therefore, an upgraded version of Wikipedia navigation might be an online application deployed on a powerful server to execute the data pipeline fast. As a starting point, the proof-of-concept satisfies the requirements proposed in Section 3, and several packages could be reused in the advanced version, too.

5 CASE STUDY

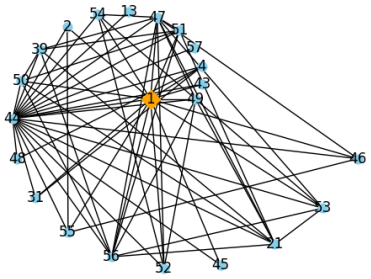
To illustrate the diverse kinds of 2D maps for Wikipedia navigation, we take the Wikipedia page "Film" as the seed, and crawl all its direct neighbors at depth one, all of which are cited on the "Film" page. Following the steps in Section 3, we set a threshold of 100 on the node degree (i.e., the number of links it has) to get a denser network



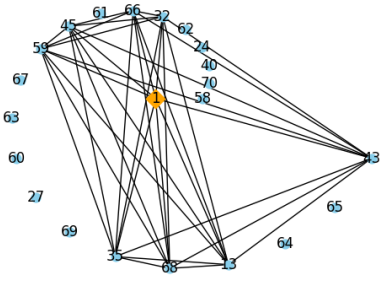
(a) Semantic content.



(b) Network structure.

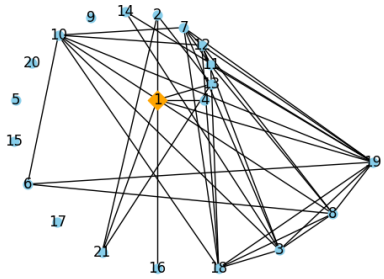


(c) Collaborative filtering.

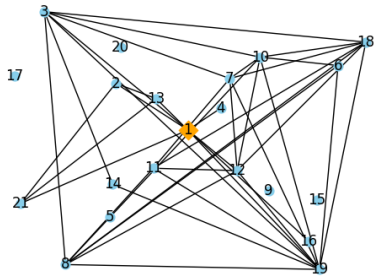


(d) A mixed PageDist [21] metric.

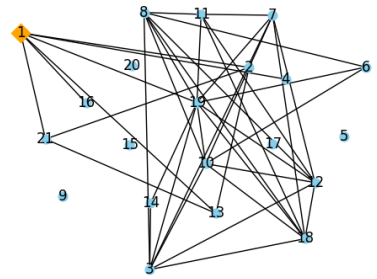
Figure 4: 2D map visualization of different sourcing methods. The distance to Node 1 is derived from feature similarity. The common layout is spiral. An orange diamond represents the root page.



(a) Spiral.



(b) Spectral.



(c) MDS.

Figure 5: 2D map visualization under different layouts. The distance to Node 1 is derived from feature similarity. The common sourcing method is semantic content. An orange diamond represents the root page.

with 8,083 directed edges and 151 nodes (pages). For a clear network visualization we only select top 20 neighboring nodes according to their feature similarity (depends on the choice of sourcing rule) to the seed node ("Film"), and display the top one-third of edges among those selected edges based on the edge weight defined by the times of concurrences on a third page.

Since we apply four kinds of sourcing and ranking methods introduced as the first group of algorithmic parameters in Section 3,

in total there are 80 nodes selected for all the maps, but some nodes might be selected by multiple sourcing rules. Table 1 shows a dictionary of them. Since an accurate user preference prediction requires real user behavior data, we choose the second algorithmic parameter introduced in Section 3 as "the distance based on feature vector similarity" instead of a user preference prediction.

As for layout options, we take the coordinates directly generated by MDS and compute spiral and spectral coordinates in a polar system, respectively. MDS exploits a pairwise distance matrix to present a sense of how near or far points are from each other in a low dimensional space (e.g., 2D plane) to users. The spiral and spectral layouts tend to prove that users may choose their personalized layouts beyond the traditional MDS visualization method, and the navigation system is flexible enough to support the function. In total, we exploit the navigation system to generate the enumerations of available sourcing-ranking methods and layouts, some of which are displayed in Figure 4 and Figure 5.

Figure 4 illustrates the spiral 2D maps of Wikipedia nodes according to four different sourcing methods. Their common algorithmic parameters suggest that the distance to the center node ("Film") corresponds to the ranking of their feature similarity to that of the Node "Film". For example, in Figure 4a, the semantic content method treats Node 4 ("Television") as the most similar neighbor to "Film", and the second one is Node 13 ("Soundtrack"). The farthest neighbor is Node 19 "Animation". In Figure 4b, according to some network science feature (e.g., degree of a node within the sandbox), the most significant two nodes are "Spectacle_(critical_theory)" and "Shooting_schedule" (a daily plan of film production). For the collaborative filtering map in Figure 4c, "Visual_effects" and "Videography" occupy the nearest two locations to the center. In the PageDist map (Figure 4d), "American_Dream" and "Bollywood" become the nearest neighbors. Users would recognize the obvious differences among the maps and choose their desired method for the following browsing.

With a limit of 20 or so nodes to a canvas in Figure 4, the four derived node sets have almost no intersection. That is, the different metrics produce very different neighborhoods in terms of their underlying node sets. If we use a larger bound of 50 nodes on a given canvas, the semantic-content set and network structure set have 10 nodes in common, the semantic-content set and collaborative filtering set share 16 nodes, while the intersection of collaborative filtering and PageDist contains 13 nodes. Going further, the first three sets (semantic, network and collaborative), have five nodes in common "Film_budgeting", "Cinematography", "Roadshow_theatrical_release", "Film_industry", "Principal_photography". The diverse navigation maps will have varying levels of utility to different user groups.

Figure 5 displays three layouts of the same subset of nodes according to the semantic content sourcing method, with a condition that the similar nodes of "Film" would be placed close to the center. For the MDS one (Figure 5c), the coordinates of all nodes are derived from a similarity matrix so that the node "Film" may not be at the center of the canvas. More importantly, MDS considers mutual similarity between any pair of nodes on the canvas, while for the other layouts, the comparative distance is only meaningful between the root node "Film" and another node.

Since only the edges with a large enough weight could be added to the map, the dense edges suggest several local clusters, such as Nodes (4, 13, 11, 12, 7), or another group (18, 3, 8, 19) in Figure 5a. Besides, the spiral layout clearly shows the similarity-based distance to the center node in an anti-clockwise order. For the second spectral layout in Figure 5b, we allocate the nodes mainly in four directions (upper right, upper left, down right, down left). It might be more difficult to compare the distance to the center for two nodes (e.g., Nodes 9 and 14), but the spectral layout makes it possible to cluster the neighboring nodes into several groups and deploy each group along a "beam". In Figure 5c, the MDS layout considers the distance matrix of all nodes in terms of the semantic vector of the corresponding Wikipedia page's title and computes their coordinates with a standard dimensionality reduction algorithm, so the root page "Film" is automatically placed in the upper left corner.

In this way, without the special color/shape, it might not attract users attention at the first glance. MDS is a popular standard visualization method, but when users choose the second algorithmic parameter about closeness in the navigator as "a probability from a predictive model", it is more difficult to define a complete distance matrix, especially between pairs of surrounding nodes.

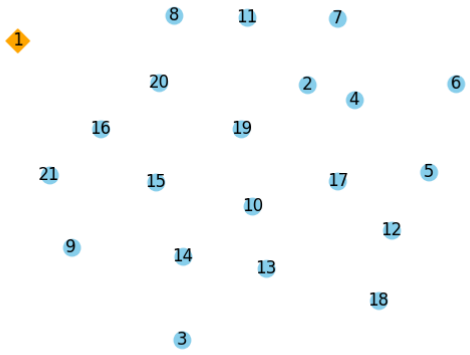


Figure 6: A non-edge version of Figure 5c with MDS.

Beyond the above algorithmic parameters and layout options, other visualization factors may be critical. Figure 6 displays the non-link version of MDS layout, in which the neighborhood is determined by a textual distance instead of link-based distance on the subnetwork. We would anticipate associating such a non-edge map with some kinds of "sliders" that would allow the picture to vary according to user feedback.

6 FUTURE WORK: DESIGN OF USER STUDY

This paper is a proof-of-concept. Before realizing a full-blown application, user study would be necessary. This is among the proposed directions of future work. Herein we outline the plan of user study to evaluate the Wikipedia navigator from two aspects: subjective feeling of transparency and general user experience; the effects of algorithmic parameters in terms of Wikipedia pages recommendation.

Questions about user experience are often an important part of a user-oriented study [6, 14, 25, 31]. We would like to design several

questions to determine the degree that users believe the navigation tool improves transparency via customizability and network visualization, as well as their preferred parameter settings. Possible data processing methods/metrics for user feedback include a correlation matrix of multiple numerical ratings [25] and statistical tests [31] to validate some significant difference.

7 CONCLUSIONS

In this paper, we have presented a proof of concept for an open navigation tool of Wikipedia pages to broaden the understanding of the information context of Wikipedia pages to a user, along with a form of algorithmic transparency for the users with the goal of enabling them to better understand why they get the current map of a vast Wikipedia network. We find that the sourcing and ranking method can greatly affect the set of finally selected nodes on the canvas, and different layouts highlight (according to the different underlying metrics) different significant neighboring nodes in the corresponding local cluster on the map.

As per the above, future work may include a more complete user study to quantify the effects of different algorithmic parameters on their preference of Wikipedia pages. In this way, we may be able to explore how the structure of Wikipedia articles influences user navigation, and improve the design of the user interface. User data could also contribute to other related research projects, such as a Wikipedia based recommender. There is also the possibility of an upgraded version of navigation tool merged into a browser (e.g., a Chrome extension) or a back-end deployment on a web server to speed up the Wikipedia page visualization.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Neukom Institute for Computational Science.

REFERENCES

[1] 2016. Research:Wikipedia Navigation Vectors. Retrieved Feb 24, 2019 from https://meta.wikimedia.org/wiki/Research:Wikipedia_Navigation_Vectors
[2] 2018. Beautiful Soup Documentation. Retrieved Jan 23, 2019 from https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3] 2018. Overview of NetworkX. Retrieved Jan 23, 2019 from https://networkx.github.io/documentation/stable/
[4] 2019. Tkinter, Python interface to Tcl/Tk. Retrieved Jan 23, 2019 from https://docs.python.org/2/library/tkinter.html
[5] 2019. Urllib, Open arbitrary resources by URL. Retrieved Jan 23, 2019 from https://docs.python.org/2/library/urllib.html
[6] Paritosh Bahirat, Yangyang He, Abhilash Menon, and Bart Knijnenburg. 2018. A data-driven approach to developing iot privacy-setting interfaces. In *23rd International Conference on Intelligent User Interfaces*. ACM, 165–176.
[7] Vevake Balaraman, Simon Razniewski, and Werner Nutt. 2018. Recoin: Relative Completeness in Wikidata. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1787–1792.
[8] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1075–1084.
[9] S Brin and L Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks And ISDN Systems*, Melissa Crouch and Tim Lindsey (Eds.). Elsevier, 107–117.
[10] Michel Buffa and Fabien Gandon. 2006. SweetWiki: semantic web enabled technologies in Wiki. In *Proceedings of the 2006 international symposium on Wikis*. ACM, 69–78.
[11] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 32–41.
[12] Trevor F Cox and Michael AA Cox. 2000. *Multidimensional scaling*. Chapman and hall/CRC.
[13] Stefano De Sabbata, Arzu Çöltekin, Kathryn Eccles, Scott Hale, and Ralph Straumann. 2015. Collaborative Visualizations for Wikipedia Critique and Activism. In *Ninth International AAAI Conference on Web and Social Media*.
[14] Ivania Donoso-Guzmán and Denis Parra. 2018. An Interactive Relevance Feedback Interface for Evidence-Based Health Care. In *23rd International Conference on Intelligent User Interfaces*. ACM, 103–114.
[15] Fabian Flöck, David Laniado, Felix Stadthaus, and Maribel Acosta. 2015. Towards Better Visual Tools for Exploring Wikipedia Article Development–The Use Case of âĂIJGamergate ControversyâĂİ. In *Ninth International AAAI Conference on Web and Social Media*. p48–55.
[16] Laxmi Amulya Gundala and Francesca Spezzano. 2018. ReadersâĂŹ Demanded Hyperlink Prediction in Wikipedia. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1805–1807.
[17] Reed H Harder, Alfredo Velasco, Michael Evans, Chuankai An, and Daniel Rockmore. 2017. Wikipedia Verification Check: A Chrome Browser Extension. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1619–1625.
[18] Christopher D. Manning Jeffrey Pennington, Richard Socher. 2014. GloVe: Global Vectors for Word Representation. Retrieved Jan 25, 2019 from https://nlp.stanford.edu/projects/glove/
[19] Daniel Lamprecht, Denis Helic, and Markus Strohmaier. 2015. Quo vadis? On the effects of WikipediaâĂŹs policies on navigation. *Links* 80 (2015), 100.
[20] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. 2017. How the structure of wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia* 23, 1 (2017), 29–50.
[21] Greg Leibon, Michael Livermore, Reed Harder, Allen Riddell, and Dan Rockmore. 2018. Bending the law: geometric tools for quantifying influence in the multinetwork of legal opinions. *Artificial Intelligence and Law* 26, 2 (2018), 145–167.
[22] G Leibon and D N Rockmore. 2013. Orienteering in Knowledge Spaces: The Hyperbolic Geometry of Wikipedia Mathematics. *PLoS ONE* (2013). https://doi.org/10.13f71/journal.pone.0067508
[23] Baiju Muthukadan. 2018. Selenium with Python. Retrieved Jan 23, 2019 from https://selenium-python.readthedocs.io/
[24] Gergely Odor, Emanuele Bugliarello, and Robert West. 2018. Poster: How Did Wikipedia Become Navigable?. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee.
[25] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 157–164.
[26] Silvan Reinhold. 2006. WikiTrails: Augmenting wiki structure for collaborative, interdisciplinary learning. In *Proceedings of the 2006 international symposium on Wikis*. ACM, 47–58.
[27] Tomás Sáez and Aidan Hogan. 2018. Automatically Generating Wikipedia Infoboxes from Wikidata. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1823–1830.
[28] Sebastian Schaffert. 2006. IkeWiki: A semantic wiki for collaborative knowledge management. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE'06. 15th IEEE International Workshops on*. IEEE, 388–396.
[29] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. 2017. Cartograph: Unlocking Thematic Cartography Through Semantic Enhancement. In *22nd International Conference on Intelligent User Interfaces, IUI 2017*. Association for Computing Machinery, 179–190.
[30] Neil Thompson and Douglas Hanley. 2018. Science Is Shaped by Wikipedia: Evidence From a Randomized Control Trial. Retrieved Jan 23, 2019 from https://ssrn.com/abstract=3039505
[31] Yuan Yao and F Maxwell Harper. 2018. Judging similarity: a user-centric study of related item recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 288–296.

Understanding the Signature of Controversial Wikipedia Articles through Motifs in Editor Revision Networks

James R. Ashford
AshfordJR@cardiff.ac.uk
School of Computer Science and Informatics, Cardiff University, UK

Liam D. Turner
TurnerL9@cardiff.ac.uk
School of Computer Science and Informatics, Cardiff University, UK

Roger M. Whitaker
WhitakerRM@cardiff.ac.uk
School of Computer Science and Informatics, Cardiff University, UK

Alun Preece
PreeceAD@cardiff.ac.uk
Crime and Security Research Institute, Cardiff University, UK

Diane Felmlee
Dhf12@psu.edu
Dept. of Sociology & Criminology, Pennsylvania State University, USA

Don Towsley
Towsley@cs.umass.edu
College of Information and Computer Sciences, University of Massachusetts Amherst, USA

ABSTRACT

Wikipedia serves as a good example of how editors collaborate to form and maintain an article. The relationship between editors, derived from their sequence of editing activity, results in a directed network structure called the revision network, that potentially holds valuable insights into editing activity. In this paper we create revision networks to assess differences between controversial and non-controversial articles, as labelled by Wikipedia. Originating from complex networks, we apply motif analysis, which determines the under or over-representation of induced sub-structures, in this case triads of editors. We analyse 21,631 Wikipedia articles in this way, and use principal component analysis to consider the relationship between their motif subgraph ratio profiles. Results show that a small number of induced triads play an important role in characterising relationships between editors, with controversial articles having a tendency to cluster. This provides useful insight into editing behaviour and interaction capturing counter-narratives, without recourse to semantic analysis. It also provides a potentially useful feature for future prediction of controversial Wikipedia articles.

KEYWORDS

Network Motifs, Wikipedia, Controversy, Complex Networks

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316754

ACM Reference Format:
James R. Ashford, Liam D. Turner, Roger M. Whitaker, Alun Preece, Diane Felmlee, and Don Towsley. 2019. Understanding the Signature of Controversial Wikipedia Articles through Motifs in Editor Revision Networks. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316754

1 INTRODUCTION

Wikipedia has become a tremendous platform for crowdsourcing knowledge, representing a cornerstone of the World Wide Web [7]. It allows the "wisdom of the crowd" [24] to potentially emerge, providing intelligence on a vast range of topics [4]. However, complex dynamics support the emergence of content, since the formation of Wikipedia articles involves both human cooperation and human conflict, based on the extent of convergent and divergent views. Narrative and counter-narrative frequently jostle for presence in a articles, representing a source of friction that is seen through editor interaction [22] and in the semantics of article content [20]. Wikipedia conveniently provides a list of controversial content that are labelled by the Wikipedia community themselves[1].

In an age of misinformation [1, 17, 26], understanding characteristics of controversial articles has increased in importance. Because of the controversial nature of some topics, the narrative in a Wikipedia article may contain misleading information that stops a neutral consensus emerging. Prior work in this area has established insights such as the predictability of controversy from editor behaviour [20], such as deletions, reversions, and statistics from the collaboration network, prediction of article quality taking insights

[1] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

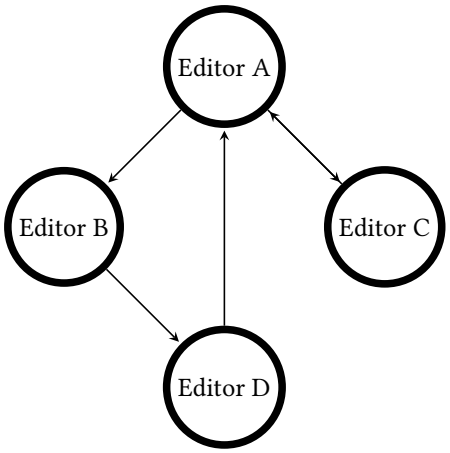


Figure 1: A network generated using the editor order A, B, D, A, C, A, from newest edit to oldest. Each editor is characterised by a letter A to D, and each occurrence in the list marks a single revision of the article by the corresponding author. An edge is formed between the current editor and its adjacent neighbour in the sequence, forming the directed edges $(A, B), (B, D), (D, A), (A, C), (C, A)$.

from multiple models [29], and interactions between users, bots, admin and pages [9]. There has also been a number of different types of network developed to assess Wikipedia articles, including collaboration networks [6] that capture the positive or negative relationship between editors, edit networks that capture "undoing" of edits by a third party [12] and affiliation networks [10].

Our focus in this paper is to further understand the relationship between small groups of editors, as induced by their editing sequences, by using a *revision network*. This does not require information on the nature of the editing undertaken - it simply captures the ordering in which editing occurs and is therefore a simple metric to infer. Editors are represented by nodes and a directed edge from node A to B indicates that "*Editor A edits the article after Editor B*" (see Figure 1). From this, we seek to determine the extent to which controversial articles exhibit a distinctive signature relative to those that are deemed non-controversial.

There has already been some consideration of revision networks in the literature [8, 11], where more recently the emphasis [27] has been to combine them with other network representations. However, given the fundamental nature of revision networks, it is interesting to question the extent to which they hold sufficient information to characterise controversial Wikipedia articles. Currently this is not well-understood, and motivates our work.

Hypotheses

We hypothesise that *interaction differences between small subgroups of Wikipedia editors is sufficient to distinguish between controversial articles and non-controversial articles.* To address this we consider the extent to which revision networks of Wikipedia articles have different local induced substructures based on their having controversial classifications. Our approach is based on techniques from complex networks [15, 16], that have been successful in classifying diverse and complex biological networks based on their latent induced subgraphs.

To achieve this at scale, and in contrast to previous literature [9, 20, 22], we assess a relatively large sample of Wikipedia articles, involving over 21,000 Wikipedia articles, by determining their subgraph ratio profiles. Each such profile represents the under and over representation of induced triads in the revision network of a Wikipedia article using 13 dimensions of connected triads, while also normalising for differences in network size.

Understanding the relationship between Wikipedia articles is important, as it allows us to determine the extent to which they may cluster. The sample we are considering represents 21,631 articles each represented in 13 dimensions. Therefore we perform dimensionality reduction, and project the subgraph ratio profiles into lower dimensional spaces. This allows us to examine the relationship both within and between controversial and non-controversial articles.

The results draw attention to distinctive clustering patterns concerning controversial Wikipedia articles. We further assess the results by examining the correlation with a range of variables, allowing to understand the role of substructures. The findings reaffirm that the sequence of editing provides an important mechanism to understand Wikipedia articles, independent of an article's topic and without recourse to semantic analysis.

2 RELATED WORK

Understanding the content of crowd-sourced platforms such as Wikipedia and the behaviour of their contributors is of wide research interest [24]. Wikipedia represents a dynamic network of articles with a structure resembling that of the World Wide Web [30], whereby dominant articles act as connectivity hubs. Dynamics also exist within the formation and maintenance of individual Wikipedia articles, through open and collaborative editing.

Interactions between editors range from positive to negative, where debates and arguments lead to different patterns of revision (e.g., [14, 20]), capturing behaviours such as vandalism [19] and the propagation of disinformation [13]. Characterising articles and contributors through revisions provides a means for Wikipedia to manage and review its

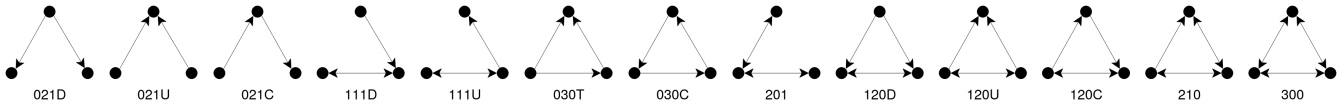


Figure 2: 13 possible combinations of connected triads in directed networks.

content. This is potentially labour intensive and has led to interest in creating and exploiting methods to detect issues (e.g.,[2, 23]).

Controversial articles have become an increasing point of focus, and characterised as such by Wikipedia. Automated methods for classifying articles have received much attention (e.g., [21, 27, 29]. The associated revision log for Wikipedia articles has been shown to provide a basis to examine potential controversy through examining the collaborative behaviour of individual editors within an article [23] or across multiple articles [28]. An article's revision log identifies the structure underlying temporal interactions [27], and provides insight into how articles and contributors' habits may evolve over time [9]. Features from the aggregation of this, such as number of edits, revision, and previous version restorations have been shown to correlate (e.g., [23]).

Treating the revision log as a network between editors [22] has been shown to provide additional useful features using graph theory and social network analysis techniques (e.g., [21]). This has ranged from global features such as the degree distribution (e.g., [23]), through to analysis of local sub-structures concerning the articles with which editors interact (e.g., [28]).

However, there has been little investigation of controversial articles based on the under or over representation of local-substructures. Called network motif analysis, this approach originated from biology [5, 15], and has been used to good effect in characterising other complex networks, including technology (e.g., [18]). In terms of Wikipedia, motif analysis has been used to determine how articles point to each other [30] and in assessing interactions between editors and different Wikipedia articles [28]. Our contribution is two-fold: firstly to use motif analysis to understand the fundamental revision networks, and secondly to consider the role of motifs in differentiating between controversial and non-controversial articles.

3 METHODS

Dataset

We collected revision history logs and article meta-data of a sample of Wikipedia articles ($N = 21,631$) through Wikipedia's web API[2]. The revision logs contain time-series events and meta-data attributing the revision to a particular user at a given time. Within this set of articles, a subset ($N = 2,661$) are considered to be 'controversial' as they are listed in Wikipedia's "List of controversial issues"[1]. The remaining articles ($N = 18,970$) are random articles that are not contained in the controversial issue list to serve as a basis for comparison. These were taken from an original sample of 23,000 articles (20,000 non-controversial and 3,000 controversial), from which articles were removed if they did not contain sufficient information for motif analysis.

Network Construction

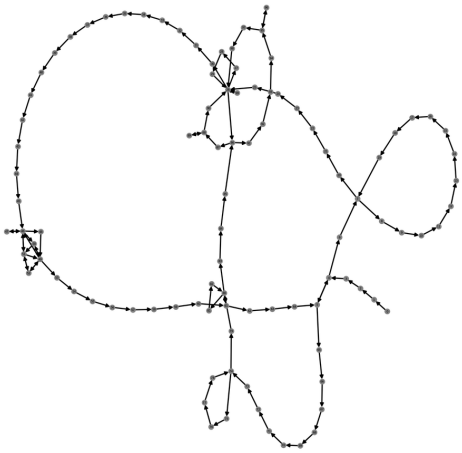
For each article, we construct a revision network (in the same manner as [11, 27]) where nodes represent unique editors and directed edges show that an editor added a revision after another editor. We traverse the revision log list to build a network that spans the article's lifetime, adding nodes and edges as they appear in each event. Specifically let the revision network of a Wikipedia article be defined by $G = (V, E)$, where each editor is represented by a node $v \in V$. An edge $(v_i, v_j) \in V$ indicates that editor v_i edits the article after editor v_j. This excludes self-loops and editor v_i editing after editor v_j multiple times does not result in a multi-edge.

A simple example can be found in Figure 1, which describes how the network is constructed. Two extreme examples from the dataset shown in Figures 3a and 3b. This form of representation is potentially useful as large articles do not typically follow a linear or incremental structure. For example, it is highly likely that users will restore work back to an earlier revision should a revision become vandalised or irrelevant. Editors are likely to refer back to previous editor's work. These behaviours result in complex sequential patterns that are captured through revision networks.

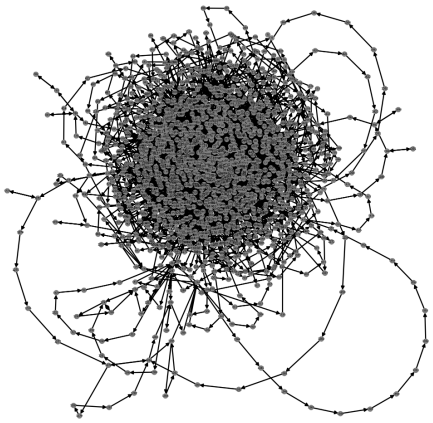
Network Motif Analysis

Network motif analysis focuses on determining the under or over representation of induced subgraphs [5, 16], as compared to an alternative sample of graphs (i.e., a null-model that acts a relevant baseline for comparison). Dyads, triads and tetrads are often considered as induced subgraphs such over and under-represented induced subgraphs are called network *motifs* or *anti-motifs*. In this analysis, we examine each article's revision network using triads, representing how all possible triples of editors may sequentially interact. Triads are sufficiently large enough to capture both direct

[2]https://en.wikipedia.org/w/api.php



(a) A revision network taken from the non-controversial *The Web Conference* Wikipedia article.



(b) A revision network taken from the controversial *Brexit* Wikipedia article. The large volume of editors produces a highly dense cluster of interactions.

Figure 3: Revision network of two articles - one non controversial and one controversial.

and indirect reciprocity between editors, while not being of a scale that is impeded by combinatorial complexity - there are 13 possible connected triads, shown in Figure 2. The coded names listed in Figure 2 are provided as part of the convention used in the triad census algorithm [3].

For each article, we calculate the subgraph ratio profile (SRP) as defined by Milo et al [15]. This accounts for variations in network size. This is achieved by determining the relative abundance of each type of triad compared to random graphs generated by the null-model. For each type of triad i, we firstly calculate Δ_i:

$$\Delta_i = \frac{(Nreal_i - < Nrand_i >)}{(Nreal_i + < Nrand_i >) + \epsilon} \quad (1)$$

where $Nreal_i$ is the number of such triads observed in the graph under observation, and $Nrand_i$ is the average frequency of such triads seen across the sample of networks composed by the null model. In this case the null model uses 100 random graphs with the same number of nodes and edges as the graph under observation. The value of ϵ is set to 4 [15, 25] to prevent the result from being misleadingly large when a subgraph rarely appears. This process is repeated for each triad i and normalised across triads to form the *subgraph ratio profile* (SRP) for a given network. The i^{th} SRP, denoted SRP_i, denotes the extent of under or over representation of the triad i, and is defined by:

$$SRP_i = \Delta_i / (\sum \Delta_i^2)^{1/2}. \quad (2)$$

Article comparison and dimensionality reduction

The SRP composed for each article provides a 13-dimensional vector whose components indicate the extent of triad representation relative to networks in the null model. To assess these collectively, we use principal component analysis (PCA) that allows the SRPs to be considered in a lower dimensional space. We apply three and two dimensional PCA. Finally, we make comparison with a number of external variables (number of editors/nodes, age of article, and revision rate) to understand potential correlations with motifs.

4 RESULTS

Subgraph Ratio Profiles

We first examine the SRPs that arise from both controversial and non-controversial articles in isolation (Figures 4 and 5 respectively). To find significant triads, we use a cutoff value of +0.3 and -0.3. We determine that controversial articles are strongly represented by triads `111D`, `111U` and `201`, which attain average *SRP* scores of 0.382, 0.375, 0.372, with relatively low dispersion (SDs of 0.136, 0.149 and 0.124 respectively). Interestingly, together these represent a chain of three nodes, where one edge is reciprocated, with the other edge covering all possible directional types (i.e., reciprocated, directed in, directed out).

In contrast, the results for the non-controversial articles provide a different profile. Here `021D` and `021U` are significantly under represented (average *SRP* scores of -0.511, -0.485), albeit with higher standard deviations present (SDs

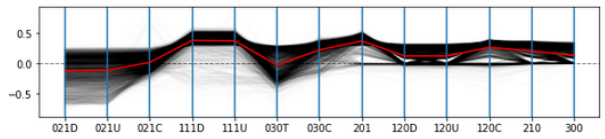


Figure 4: Subgraph ratio profiles of all controversial articles. Average displayed in red.

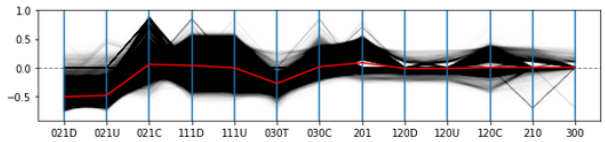


Figure 5: Subgraph ratio profiles of all non-controversial articles. Average displayed in red.

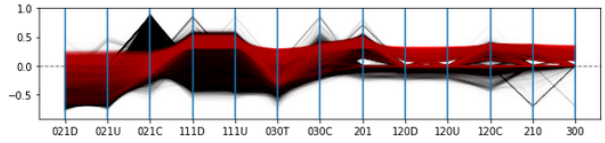


Figure 6: Subgraph ratio profiles of all articles, with controversial articles displayed in red and non-controversial in black.

of 0.192, 0.2). Interestingly these anti-motifs (021D and 021U) relate to a lack of subgraphs where directed edges either emanate from or are received by a single node in the triad. Such configurations relate to the role of a mediating editor that may be presented with or respond to the editing of others. In other words, such mediators have a reduced role in non-controversial articles.

The comparison between these subgraph ratio profiles is shown in Figure 6. These profiles are quite distinct. We also calculate the Pearson correlation coefficient for each distinct pair of articles in three groups - controversial articles, non-controversial articles and all articles. Controversial articles provide the greatest correlation to each other (M=0.41193, SD=0.41262). Non-controversial articles have a lower mean correlation (M=0.37498, SD=0.36749) which is similar to the result when considering all articles together (M=0.37569, SD=0.36844).

Principal Component Analysis

We analyse the 21,631 subgraph ratio profiles to determine the relationship in terms of relative clustering. We apply PCA in order to reduce the 13 dimensions of the SRPs down to a more manageable form. We initially project the SRPs into 3-dimensional space for clarity, as seen in Figure 7. This presents a distinctive region where controversial articles are dominant. This provides evidence for a distinction between

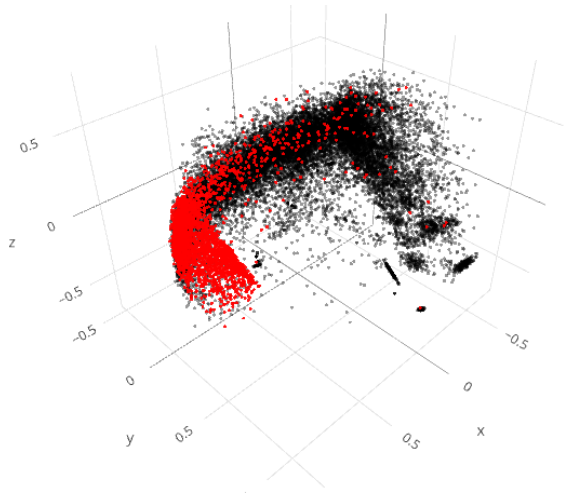


Figure 7: PCA scatter plot repression of the 13-point feature vector in 3D clustering space.

the controversial and non-controversial articles, consistent with the variation in motifs identified in the previous section.

The PCA coefficients (Table 1) that define the three dimensions reveals that principal component one (x axis in Figure 7) primarily depends on triads 111D and 111U. Principal component two (y axis in Figure 7) primarily depends on triad 021C. The third principal component primarily depends on triads 021C and 030C. However, we also note that when represented in the three dimensional space, the revision networks have limited dispersion in the third dimension (i.e., vertical dimension as plotted).

Calculating the percentage of explained variance by principal component confirms that the first principal component produces 53.7% of the shared variance, the second produces 22.3% and the third produces the least with 6.7%. This confirms that the third principal component provides a limited contribution to representation of the total variance across the significant ratio profiles. This supports representation through two principal components, as plotted in Figure 8, with the relative composition of each principal component being near identical to PC-1 and PC-2 in Table 1. As anticipated, this is similar dependency on the first and second primary components in the three dimensional representation.

Representation in two dimensions further clarifies the distinction between controversial and non-controversial revision networks. In particular, from Figure 8 we note that both classes of article exhibit a similar maximum and minimum range against principal component two, which is primarily defined by the linear path between three nodes (021C). However, it is the variation in the first principal component, dominated by 111D and 111U, which represent linear paths

Table 1: PCA coefficients displaying the strongest triads

	021D	021U	021C	111D	111U	030T	030C	201	120D	120U	120C	210	300
PC-1	0.332	0.321	-0.126	0.509	0.522	0.145	0.205	0.318	0.078	0.079	0.204	0.117	0.065
PC-2	0.427	0.438	0.643	-0.278	-0.242	0.207	0.104	-0.052	0.059	0.06	0.047	0.074	0.06
PC-3	0.074	0.127	-0.609	-0.398	-0.332	0.136	0.441	0.115	0.131	0.130	0.19	0.167	0.112

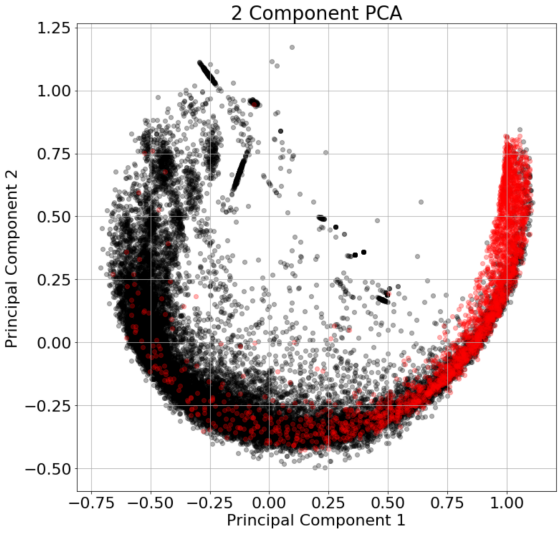


Figure 8: PCA scatter plot repression of the 13-point feature vectors in 2D clustering space.

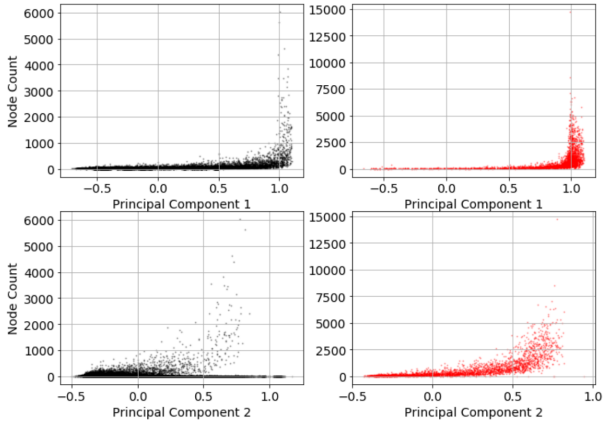


Figure 9: Scatter plot of each principal component combined with node count for both non-controversial and controversial articles.

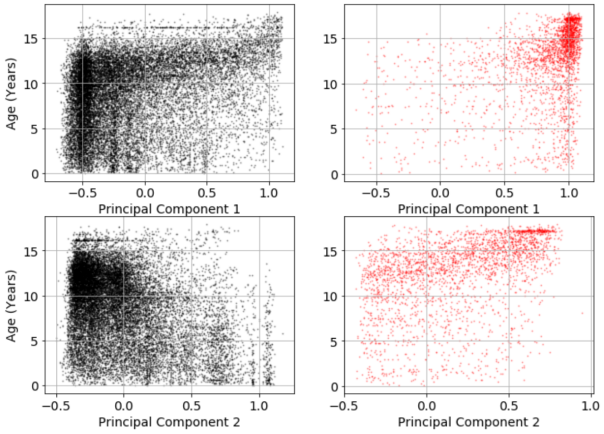


Figure 10: Scatter plot of each principal component combined with article age for both non-controversial and controversial articles.

with reciprocation on one edge, that differentiate the non-controversial from controversial. High values in principal component one correlate with controversial articles - in other words, controversial articles exhibit more reciprocation on top of possible linear paths.

Additional Metrics

We examine the relationship between revision networks and primary external variables (number of editors/nodes, age of article, and revision rate) using motifs. Specifically, using the dimensions of two-dimensional PCA analysis, we examine the correlation with the external variables, and how this differs between controversial and non-controversial articles. The results are shown in Figures 9, 10 and 11.

The greatest differences between controversial and non-controversial articles occur with respect to article age (Figure 10). Here, controversial articles with high age cluster against high values of principal component one, and to some extent this occurs for principal component two. This contrasts against the clustering seen for non-controversial articles.

Figure 10 also shows that controversial articles have a tendency to be older. The high values in principal component one and two which align with dense clustering of controversial articles show that while such articles accumulate the linear revision path between authors (`021C` which dominates principal component two), controversial articles also accumulate instances of linear paths where one edge is reciprocated (i.e., `111D` and `111U` which dominate principal component one).

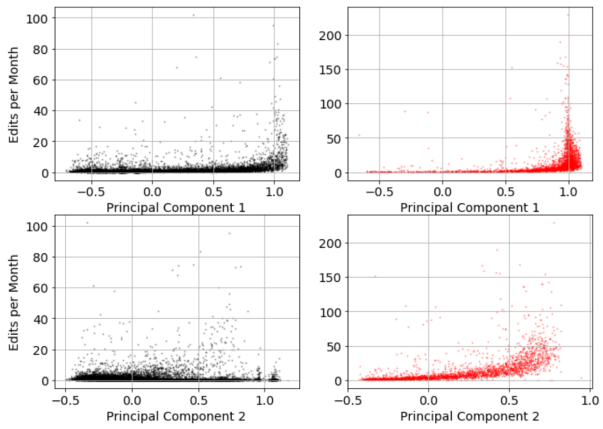


Figure 11: Scatter plot of each principal component combined with edit rate (mean number of edits per month) for both non-controversial and controversial articles.

5 DISCUSSION

Motif analysis of revision networks gives insight into how the temporal editing relationship between small groups of Wikipedia authors create signatures that allow controversial articles to be distinguished. In contrast to previous work, we have investigated this using a relatively large sample of Wikipedia articles, where distinct patterns emerge. This provides strong support for our hypothesis, and reaffirms the importance of the revision network as a simple but fundamental element in editing Wikipedia.

Through motif analysis, we identify that reciprocation on linear paths among triads of editors in the revision network is over represented in controversial Wikipedia articles. These motifs are defined by the triads `111D`, `111U` and `201`. In contrast, the revision networks from non-controversial articles exhibit two anti-motifs, involving the under representation of triads involving two directed edges either arriving at or emanating from a mediating node (triads `021D` and `021U`). These motifs and the underlying subgraph ratio profiles represent an unusual and distinctive profile that we believe represent distinctive "super-families" beyond those seen in other technologically related networks, such as the world wide web [30].

Performing dimensionality reduction upon the subgraph ratio profiles from each revision network allows us to further understand the relationships between Wikipedia articles. Our analysis shows that the structure of the data is amenable to reduction to two dimensions, where the principal components are dominated by triads `111D` and `111U` in the first component, and mainly `021C` in the second component, but with lesser contributions from triads `021D` and `021U`.

The results from two-dimensional principal analysis are illuminating - the dominant triads in both components, as listed above, correspond to linear paths, i.e., open triads which represent sequences of editing without indirect reciprocity. The extent and format of reciprocated (i.e., bidirectional) edges on these open triads is sufficient to define the two principal components. The dominant triads in the first principal component each involve reciprocation on one edge, where as interestingly, in the second principal component, the dominant triads are open triads with no reciprocated edges. From this we deduce that short paths, rather than short loops of editing that represent indirect reciprocity, are important features in characterising Wikipedia revision networks.

We also observe through two-dimensional principal component analysis that it is the first principal component that strongly distinguishes between the revision networks of controversial and non-controversial articles. The dominant triads defining this capture the extent of direct reciprocation being present. Finally, through consideration of the principal components against additional external variables, we find in particular that article age plays a role in distinguishing the controversial articles. High values in both principal components aligns with strongest clustering of controversial articles, which is not the case for non-controversial articles.

6 CONCLUSION

The analysis has given insights into the structure underlying revision networks from Wikipedia articles, and has shown that a relatively small number of features, in terms of substructures in revision networks, characterise controversial Wikipedia articles. The results have identified key clusters of editorial interactions to this effect, in support of the hypothesis. These are distinctive and indicate that the revision networks for controversial and non-controversial Wikipedia articles have differentiated subgraph ratio profiles. Our study gives understanding as to how prediction or classification of articles can be enhanced using the latent structures relating to editor behaviour. This also reaffirms the importance of the revision network as a simple but useful representation for assessment of Wikipedia articles.

ACKNOWLEDGEMENTS

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. Spread of (mis)information in social networks. *Games and Economic Behavior* 70, 2 (2010), 194–227.

[2] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 277–288.

[3] Vladimir Batagelj and Andrej Mrvar. 2001. A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social Networks* 23, 3 (2001), 237 – 243. https://doi.org/10.1016/S0378-8733(01)00035-1

[4] Daren C Brabham. 2008. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence* 14, 1 (2008), 75–90.

[5] Dave Braines, Diane Felmlee, Don Towsley, Kun Tu, Roger M Whitaker, and Liam D Turner. 2018. The role of motifs in understanding behavior in social and engineered networks. In *Next-Generation Analyst VI*, Vol. 10653. International Society for Optics and Photonics, 106530W.

[6] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise Van Raaij. 2009. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th international conference on World wide web*. ACM, 731–740.

[7] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing Systems on the World-Wide Web. *Commun. ACM* 54, 4 (April 2011), 86–96.

[8] Takashi Iba, Keiichi Nemoto, Bernd Peters, and Peter A Gloor. 2010. Analyzing the creative editing behavior of Wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences* 2, 4 (2010), 6441–6456.

[9] David Jurgens and Tsai-Ching Lu. 2012. Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia.

[10] Gerald C Kane and Sam Ransbotham. 2016. Research noteâĂŤcontent and collaboration: an affiliation network approach to information quality in online peer production communities. *Information Systems Research* 27, 2 (2016), 424–439.

[11] Brian Keegan, Darren Gergle, and Noshir Contractor. 2012. Staying in the loop: structure and dynamics of Wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM, 1.

[12] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 453–462.

[13] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 591–602.

[14] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the wikipedians talk: Network and tree structure of wikipedia discussion pages.. In *Sixth International AAAI Conference on Weblogs and Social Media*. 177–184.

[15] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. Superfamilies of evolved and designed networks. *Science* 303, 5663 (2004), 1538–1542.

[16] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[17] Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. 2015. Collective attention in the age of (mis)information. *Computers in Human Behavior* 51 (2015), 1198–1204.

[18] Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. 2012. Network Analysis of Recurring YouTube Spam Campaigns. (2012). arXiv:1201.3783

[19] Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic vandalism detection in Wikipedia. In *European conference on information retrieval*. Springer, 663–668.

[20] Hoda Sepehri Rad and Denilson Barbosa. 2012. Identifying Controversial Articles in Wikipedia: A Comparative Study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym '12)*. ACM, New York, NY, USA, Article 7, 10 pages.

[21] Hoda Sepehri Rad and Denilson Barbosa. 2012. Identifying controversial articles in Wikipedia: A comparative study. In *Proceedings of the eighth annual international symposium on wikis and open collaboration*. ACM, 7.

[22] Hoda Sepehri-Rad and Denilson Barbosa. 2015. Identifying Controversial Wikipedia Articles Using Editor Collaboration Networks. *ACM Trans. Intell. Syst. Technol.* 6, 1, Article 5 (March 2015), 24 pages.

[23] Hoda Sepehri-Rad and Denilson Barbosa. 2015. Identifying controversial Wikipedia articles using editor collaboration networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 1 (2015), 5.

[24] James Surowiecki and James. 2005. *The wisdom of crowds*. Anchor Books. 306 pages. https://dl.acm.org/citation.cfm?id=1095645

[25] Kun Tu, Jian Li, Don Towsley, Dave Braines, and Liam D. Turner. 2018. Network classification in temporal networks using motifs. In *Proceedings of the 3rd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*.

[26] M Del Vicario, A Bessi, F Zollo, F Petroni, A Scala, and G Caldarelli. 2016. The spreading of misinformation online. *Proc National Academy of Sciences* 113, 3 (2016).

[27] Guangyu Wu and Pádraig Cunningham. 2015. Integration of multiple network views in Wikipedia. *Knowledge and Information Systems* 45, 2 (2015), 473–490.

[28] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. 2011. Characterizing Wikipedia pages using edit network motif profiles. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 45–52.

[29] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. 2012. Classifying Wikipedia articles using network motif counts and ratios. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12*.

[30] Vinko Zlatić, Miran Božičević, Hrvoje Štefančić, and Mladen Domazet. 2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E* 74, 1 (2006), 016115.

A Graph-Structured Dataset for Wikipedia Research

Nicolas Aspert
LTS2, EPFL, Station 11, CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

Volodymyr Miz
LTS2, EPFL, Station 11, CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

Benjamin Ricaud
LTS2, EPFL, Station 11, CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

Pierre Vandergheynst
LTS2, EPFL, Station 11, CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

ABSTRACT

Wikipedia is a rich and invaluable source of information. Its central place on the Web makes it a particularly interesting object of study for scientists. Researchers from different domains used various complex datasets related to Wikipedia to study language, social behavior, knowledge organization, and network theory. While being a scientific treasure, the large size of the dataset hinders preprocessing and may be a challenging obstacle for potential new studies. This issue is particularly acute in scientific domains where researchers may not be technically and data processing savvy. On one hand, the size of Wikipedia dumps is large. It makes the parsing and extraction of relevant information cumbersome. On the other hand, the API is straightforward to use but restricted to a relatively small number of requests. The middle ground is at the mesoscopic scale, when researchers need a subset of Wikipedia ranging from thousands to hundreds of thousands of pages but there exists no efficient solution at this scale.

In this work, we propose an efficient data structure to make requests and access subnetworks of Wikipedia pages and categories. We provide convenient tools for accessing and filtering viewership statistics or "pagecounts" of Wikipedia web pages. The dataset organization leverages principles of graph databases that allows rapid and intuitive access to subgraphs of Wikipedia articles and categories. The dataset and deployment guidelines are available on the LTS2 website https://lts2.epfl.ch/Datasets/Wikipedia/.

KEYWORDS

Dataset, Graph, Wikipedia, Temporal Network, Web Logs

ACM Reference Format:
Nicolas Aspert, Volodymyr Miz, Benjamin Ricaud, and Pierre Vandergheynst. 2019. A Graph-Structured Dataset for Wikipedia Research. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316757

1 INTRODUCTION

Wikipedia is one of the most visited websites in the world. Millions of people use it every day searching for answers to various questions ranging from biographies of popular figures to definitions of complex scientific concepts. As any other website on the Web, Wikipedia stores web logs that contain viewership statistics of every page. Worldwide popularity of this free encyclopedia makes these records an invaluable resource of data for the research community. In this work, we present a convenient and intuitive graph-based toolset for researchers that will ease the access to this data and its further analysis.

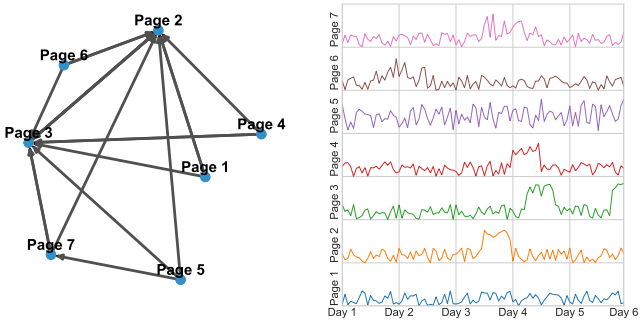


Figure 1: A subset of Wikipedia web pages with viewership activity (pagecounts). Left: Wikipedia hyperlinks network, where nodes correspond to Wikipedia articles and edges represent hyperlinks between the articles. Right: hourly page-view statistics of Wikipedia articles (right).

Wikimedia Foundation, the organization that hosts Wikipedia, makes the web activity records and the hyperlinks structure of Wikipedia publicly available so anyone can access the records either through an API or through the database dump files. Even though the data is well structured, efficient pre-processing and wrangling requires data engineering skills. First, the dumps are very large and it takes a long time for researchers to load and filter them to get what they need to study a particular question. Second, although the API is well documented and easy to use, the number of queries and the response size are very limited.

Even though the API is quite convenient, it can cause reproducibility issues. The network of hyperlinks evolves with time so the API can only provide the latest network configuration. To solve this problem, as a workaround, researchers use static preprocessed datasets. Two of the most popular datasets for Wikipedia network research are available on the SNAP archive, Wikipedia Network of Hyperlinks [11] and Wikipedia Network of Top Categories [6, 10, 16]. The initial publications referring to these datasets have been cited more than 1000 times, showing the high interest in these datasets. These archives were created from Wikipedia dumps

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316757

in 2011 and 2013 respectively. However, Wikipedia has evolved since then and Wikipedia research community would benefit from being able to access more recent data.

Multiple studies have analyzed Wikipedia from a network science perspective and have used its network structure to improve Wikipedia itself or to gain insights into collective behavior of its users. In [17], Zesch and Gurevych used Wikipedia category graph as a natural language processing resource. Buriol et al. [4] studied the temporal evolution of Wikipedia hyperlinks graph. Bellomi and Bonato conducted a study [3] of macro-structure of English Wikipedia network and cultural biases related to specific topics. West et al. proposed an approach enabling the identification of missing hyperlinks in Wikipedia to improve the navigation experience [12].

Another direction of Wikipedia research focuses on the pagecounts analysis. Moat et al. [9] used Wikipedia viewership statistics to gain insights into stock markets. Yasseri et al. [15] studied editorial wars in Wikipedia analyzing activity patterns in viewership dynamics of articles that describe controversial topics. Mestyán et al. [7] demonstrated that Wikipedia pagecounts can be used to predict the popularity of a movie. Collective memory phenomenon was studied in [5], where authors analyzed visitors activity to evaluate the reaction of Wikipedia users on aircraft incidents.

The hyperlink network structure, on one hand, and the viewership statistics (pagecounts) of Wikipedia articles, on the other hand, have attracted significant attention from the research community. Recent studies open new directions where these two datasets are combined. The emerging field of spatio-temporal data mining [2] highlighted an increasing interest and a need for reproducible network datasets that contain dynamically changing components. Miz et al [8] adopted an anomaly detection approach on graphs to analyze the visitors' activity in relation to real-world events.

Following the recent advances of scientific research on Wikipedia, in this work, we focus on two components: the *spatial* component (Wikipedia hyperlinks network) and the *temporal* component (pagecounts). We design a database that allows querying this hybrid data structure conveniently (see Fig. 1). Since Wikipedia web logs are continuously updating, we designed this database in a way that will make its maintenance as easy and fast as possible.

2 DATASET

There are multiple ways to access Wikipedia data but none of them provide native support of a graph data structure. Therefore, if researchers want to study Wikipedia from the network science perspective, they have to create the graph themselves, which is usually very time-consuming. To do that, they need to pre-process large dumps of data or to use the limited API.

In spatio-temporal data mining [2], researchers are most interested in the dynamics of the networks. Hence, when it comes to Wikipedia analysis, one needs to merge the hyperlinks network with page-view statistics of the web pages. This is another large chunk of data, which requires another round of time-consuming pre-processing.

After the pre-processing and the merge is completed, researchers usually realize that they do not need the full network and the entire history of visitors' activity. However, there is no easy workaround: in order to get a certain subset of pages for a specified period, everyone has to perform the aforementioned steps.

In this paper, we propose a graph-based solution that eliminates the pre-processing steps described above. We present a graph database that simplifies access to Wikipedia data dumps and its viewership statistics. With a set of intuitive queries, we provide the following features:

- Get relatively large subgraphs of Wikipedia pages (1K–100K nodes) without redirects.
- Use filters by the number of page views, category/sub-category, graph measures (n-hop neighborhood of a node, node degree, page rank, centrality measures, and others).
- Get viewership statistics for a subset/subgraph of Wikipedia pages.
- Get a subgraph of pages with a number of visits higher than a threshold, in a predefined range of dates.

The database allows its users to return subgraphs with millions of links. However, requesting a large subgraph from the database may take several hours. Besides, it may require a large amount of memory on the hosting server. Such queries may cause an overload of the database server that has to process queries from multiple users at the same time. Therefore, instead of setting up a remote database server, we have decided to provide the code to deploy a local or cloud-based one from Wikipedia dumps. This will allow researchers to explore the dataset on their own server, create new queries and, possibly, contribute to the project.

Lastly, the database will be updated every month and will be consistent with the latest Wikipedia dumps. This gives the researchers the ability to reproduce previous studies on Wikipedia data and to conduct new experiments on the latest data. The dataset and the deployment instructions are available online [1].

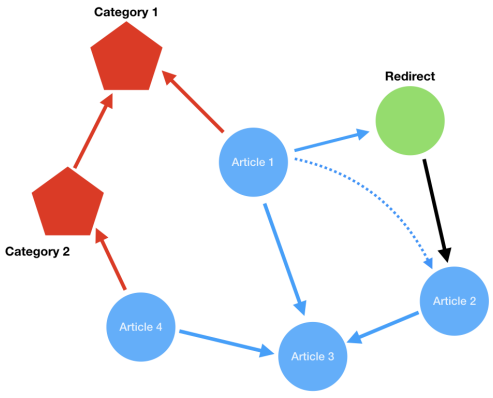


Figure 2: Wikipedia graph structure. In blue: articles and hyperlinks referring to them. In red: category pages and hyperlinks connecting the pages or subcategories to parent categories. In green: a redirected article, i.e. Article 1 refers to Article 2 via the redirected page. In black: a redirection link. The blue, dashed line, is the new link created from the redirection.

Table 1: Entities in the graph

Name	Nature	Description
article	node	Wikipedia article
category	node	Wikipedia category article
links_to	link	hyperlink between 2 articles
belongs_to	link	hyperlink between an article or subcategory and a category page

3 FRAMEWORK

3.1 Graph structure

Wikipedia network of articles is stored in a property graph database. The graph is a multigraph with different kinds of nodes and links. The different objects are described in Table 1 and on Fig. 2.

Wikipedia articles are nodes of the graph and the hyperlinks pointing to different pages are recorded as directed edges between the nodes. The categories of Wikipedia articles are structured as a graph as well. Indeed, in the encyclopedia, categories are pages with a textual description, and they refer to their elements (articles or subcategories) with hyperlinks. Hyperlinks are present in each article pointing to the categories they belong to. Inside the graph database, articles and category pages are distinguished as nodes of different nature (different labels).

In order to navigate between articles and categories conveniently, we introduce two types of links. The "links_to" relations are hyperlinks between articles (excluding categories), and the "belongs_to" relations are linking articles to their categories or subcategories to their parent categories. These latter edges are built from the hyperlinks within pages pointing to category pages.

Graph structure. The category structure is shaped as a tree. The advantage of such structure is that it is easy to handle it when the user creates articles and wants to classify them in subcategories. However, it makes it difficult to retrieve the set of all articles belonging to a given category (or subcategory). One has to explore all the hierarchy of subcategories within it and collect all the encountered articles. Therefore, we choose the graph database structure to simplify this task. Traversing and performing the breadth-first search in the graph is one of the basic functions of a graph database making this solution a more efficient alternative.

Redirects. In order to handle renamed or merged pages, Wikipedia relies on a redirection approach. When renaming a page, moderators create a new page with a new title but they do not remove the initial page in order for the hyperlinks from articles pointing to remain valid. Instead, in order to avoid this, the initial page becomes a "redirect", a page that automatically redirects a visitor to a new page. Redirect pages are invisible to users. We removed these redirection pages from our dataset by redirecting the hyperlinks pointing to the correct article (the blue dashed arrow of Fig. 2). First, it simplifies the queries when exploring the graph. Second, it makes it easier for users to understand the structure. Lastly, it halves the number of nodes in the graph: at the time of writing, the number of articles in the English Wikipedia is close to 6 million while the number of redirects is around 8 million.

3.2 Time series of visits

The time series of visits are stored separately in a NoSQL database in the form of a collection of indexed *key:value* pairs. Each key is a pair *(page id, time-stamp)* and the value is the *number of visits* during the hour given by the time-stamp for the page associated to the page id.

This structure provides a flexible way of recording new entries following the evolution of time, the page creation and deletion that occur in the encyclopedia. Querying a specific period of time is very convenient and efficient as well. It is done by submitting a request with a specific range of key values (a range applied to the time-stamp key of the key couple).

In order to reduce the amount of data to be stored, we introduce a threshold for a number of visits per page per day. We store the number of hourly visits for an article if the daily total of its visits is above this threshold (100 by default). This reduces the number of entries by an order of magnitude without losing relevant information. The database handles missing records automatically, which is also very convenient.

3.3 Data extraction and pre-processing

Before creating the database, we perform the following pre-processing steps. After having downloaded Wikipedia dumps [13], we parse the SQL files to extract the titles of articles and categories, page and category ids, and the hyperlinks. Before storing the data in the graph database, we remove the redirects and modify the hyperlinks pointing to them to link to the correct articles. After these steps are completed, we load the data into the graph database.

We download the pagecounts dumps [14] (number of visits per page per hour), and extract the hourly visits. As stated in section 3.2, we remove entries with a low number of visits. If a page has less than 100 daily visits, we do not store visit records for that page and that day. We store the values above this daily threshold in the NoSQL database with an hourly resolution.

3.4 Database update

The separation of the graph and time series data simplifies the update process and the maintenance. We update every part separately and in a different manner.

We update the graph database periodically. At the time of writing, we do it on a monthly basis. The update follows monthly releases of Wikipedia dumps. Every month, we compare the new and the previous version dumps of pages and links. We add the new nodes and links to the database and delete the removed ones.

We perform daily updates of the time series database. Every day, we add 24 new entries (one per hour) for each article (only those entries that surpass the 100 daily visits threshold).

4 PERFORMANCE OF THE QUERIES

We constructed the graph of English Wikipedia pages based on the August 1st 2018 SQL dumps. After resolving redirects, the graph consists of about ca. 7.4 million articles, comprising both regular pages (ca. 5.7 million) and category pages (ca. 1.7 million), and ca. 511 million edges. Once the data was imported into the graph database, we ran queries to extract various sub-graphs, e.g. retrieve all pages and subcategories belonging to a given category and all the links between these pages. Table 3 demonstrates query results and the time required to process them. The presented results have been computed on a 24 cores Intel Xeon E5 system, equipped with SSD drives and using the Neo4j open-source database. Given the highly connected structure of Wikipedia, we had to restrict the depth of certain queries, as the returned set expands dramatically.

While it is possible to use traditional relational databases to store the pages and links information, retrieving a subgraph using such a structure would require an increasing number of subqueries when increasing the depth of the subgraph queried, resulting in longer processing time and complex query syntax. The subgraph requires multiple queries to find all the nodes belonging to the subgraph, then an additional search to find all the edges connecting any of the nodes in the set. Experiments have been conducted using direct processing of pre-processed page and link data using Apache Spark, and also using a relational database (PostgreSQL). Data used to perform the databases comparison is based on a trimmed down version of the Wikipedia SQL dumps, with additional processing to replace the target of links, expressed as page title and namespace, by the page unique id, and removing redirects. This preprocessing, combined with database indexes, leads to simpler and faster
queries.

Queries performed on a database are often at least an order of magnitude faster than direct queries on the raw page and links data using Apache Spark. For instance, querying the sub-category graph of the "Physics" category with a depth 2 requires approximately 5 minutes to retrieve all the nodes belonging to the subgraph, and also several additional minutes to retrieve its edges, whereas the same data is completely extracted in less than 5 seconds using the graph database. Using a relational database improves the situation, as the nodes of the subgraph are returned in less than a second. Returning the edges from the subgraph remains however time-consuming (ca. 10 to 40 seconds in our experiments), in addition to requiring multiple nested queries whose complexity increases with the search depth. In that particular example, given the relatively small size of the result, timings can be heavily impacted by the cache of each application, as well as by the system they run on, especially by the presence of SSD vs. HDD. For instance, using a database query (on the graph database or the relational one) to retrieve a subgraph of depth 3, then retrieving the same subgraph of depth 2 will most likely only use the cache and yield much faster results. When the search depth increases sufficiently, the relational database can lead to faster processing than the graph database, at the expense of query complexity.

Table 2: Number of nodes in the subgraph of the "Physics" category

Depth	Articles	Subcategories
1	69	27
2	2'263	206
3	10'128	970
4	33'917	3'711
5	80'349	16'917
6	232'818	74'004
7	2'041'232	251'551

Similarly, we queried subgraphs consisting of page neighbors (i.e. connected via a "links_to" relation), up to a certain depth. We also restricted the queries by the number of outgoing links from the top page since some of them have a huge number of direct connections. We provide the results of these queries in Table 4. Increasing the depth of such queries (e.g. for depth greater than one) leads to the large responses, resulting in long processing time (cf. the "Computer science" entry in Table 4).

5 USE CASES AND APPLICATIONS

5.1 Subgraphs of categories

Wikipedia articles are classified according to the category hierarchy established by the contributors. The absence of strict guidelines or strong authority on the category labeling has led to a complex category schema. Gathering all the pages belonging to a category is a difficult task at the moment. It requires visiting all the subcategories belonging to the initial category and collecting the articles they refer to. Furthermore, the absence of dedicated curation leads to the collection of several subcategories (and hence articles) only remotely related to the original one. Those sub-categories can be very generic and encompass a large number of articles, e.g. one of the sub-categories of "Physics" is "Writing systems" (linked via "Physical systems"). Indeed, the very deep hierarchy and the lack of tools for accessing the network of categories make it impossible to have a global view on the structure and efficient maintenance.

To illustrate the complexity of the category structure, we run multiple different queries. Each query defines a category and asks for all the articles belonging to it and its subcategories. The results are shown in Table 3. In the case of broad categories, the number of articles grows rapidly as we go deeper in the subcategory hierarchy. Each subcategory may have subcategories of its own and we define the depth to be the distance in hops from the initial category to the furthest subcategory in the subcategory tree. For instance, the category *Physics* already contains 33'917 articles and 972'206 hyperlinks at depth 4 and more than 2 million pages when articles are collected up to subcategory depth 7 as shown in Table 2. This is one-third of the articles of the English Wikipedia. This result is surprisingly large and additional investigation is required to understand the structure and check its correctness. The network is so connected that the number of links and the time to retrieve the data grows very quickly. After 4 hops in the category tree, it reaches

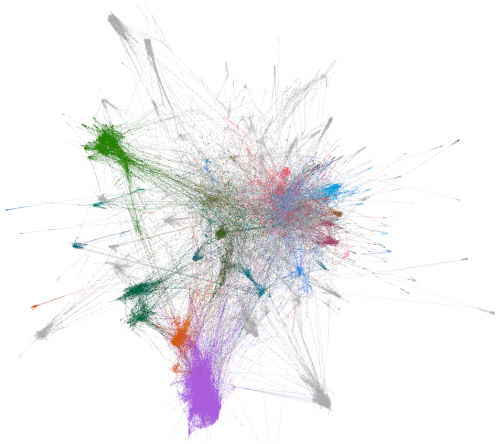


Figure 3: Network view of a reduced subset of Wikipedia web pages (~20K nodes, ~100K edges) using the method described in [8]. Nodes correspond to popular articles with spikes of visits during the period Oct. 2014 - Apr. 2015. They are connected by links with strength related to the correlation of their viewership activity. Real-world events trigger spikes in the number of visits in groups of pages, forming the clusters in the network.

tens of thousands of pages and more than a million links for some categories.

This complexity in the category hierarchy makes it memory-expensive to query subgraphs of articles in the same category. Even though our solution slows down when the network expands, it is possible to query these subgraphs. Hence, our proposed database opens new avenues to the popularization of research on large sub-networks of categories. This may give a better understanding of the category and article structures. The results may lead to a better organization of categories and a more efficient process of verification of consistency in them.

5.2 Combining the hyperlink graph and time series of visits

It was shown that real-world events can be detected and tracked using Wikipedia viewership data. Besides, it is possible to use this data to detect abnormal patterns of visits in groups of connected Wikipedia articles [8]. For example, popular sports events such as Super Bowl, NBA playoffs, and FIFA World Cup, can be detected just by looking at Wikipedia viewership dynamics and the hyperlinks structure. Moreover, dramatic events such as airplane crashes or terrorist attacks can be spotted and analyzed. Also, the authors showed that it is possible to gain interesting insights related to the history of an event and its popularity among the users of the Web. This dataset will allow further investigations in this direction.

Fig. 3 illustrates the result of the anomaly detection algorithm presented in [8]. Using the combination of the hyperlink network and the visitors' activity, the authors detected the groups of articles with simultaneous spikes in viewership dynamics. Besides, the authors used the dataset presented in this paper. They have published an interactive version of the results online as a part of the Wiki Insights project [1].

This approach allows selecting a subset of important hyperlinks out of the large amount present in each Wikipedia article. Indeed, following all the hyperlinks of some selected pages leads to very large graphs, as shown in Fig. 4. New approaches are needed in order to create meaningful subgraphs by following only a reduced number of hyperlinks. This will help extract or emphasize particular types of information present in the network. In order to perform such studies, researchers can define different filters when querying the database we propose.

6 CONCLUSION AND FUTURE DEVELOPMENT

In this paper, we presented a graph database to store and access Wikipedia web network and viewership activity of the pages. The main goal of this project was to provide a convenient tool for researchers working on Wikipedia and analyzing dynamic properties of this network. We designed the database with the idea of reproducible research in mind. We want this project to become an important building block in Wikipedia research community that should speed up the research process.

6.1 Reproducible research

In science, it should be mandatory for any published results to be reproducible. This implies an unlimited access to the data used for the experiments. However, when the dataset evolves with time, as it is in the case with Wikipedia articles and viewership statistics, it may be difficult to recover the exact data used in a given study. Some articles may have been removed or some links may have appeared after the publication of a scientific work. A database designed for scientists must include a mechanism for allowing experiments to be reproducible. One of the simplest solutions is as follows. The graph and time series are saved (frozen) every month. This allows the retrieval of graphs and pagecounts for any period in the past, while reducing the amount of memory required by the database.

6.2 Potential benefits for Wikipedia

A better understanding of Wikipedia structure, both from an article and a category point of view, is an important matter for the encyclopedia and the organization of its knowledge. Finding missing hyperlinks, suggesting links on pages creation, monitoring Wikipedia visitors activity, or structuring the category tree, are among the numerous possible applications.

6.3 Enriching the database

There are multiple ways for improving and enlarging the dataset. At the moment it contains only English Wikipedia. The number of languages can be extended. Articles with the same topic in different languages are naturally linked together in Wikipedia, which perfectly fits into the graph database framework. One could think of adding more data such as the text of articles for instance. To limit the graph database size, the best way would be to have a convenient

[1] https://wiki-insights.epfl.ch/

Table 3: Size and performances for different subgraph requests

Category	Articles	Hyperlinks	Subcategories	Search depth	Processing time
Philosophy	571	5'165	202	2	0.4 s
	5'370	177'754	1'144	3	29.7 s
	26'480	1'094'550	4'084	4	574 s
Physics	2'263	27'911	207	2	3.3 s
	10'128	223'870	971	3	55 s
	33'917	972'206	3'712	4	501 s
Science	1'762	19'189	455	2	3 s
	18'751	260'043	2'842	3	292 s
Actors	1'107	3'313	654	2	1.6 s
	10'805	47'196	2'922	3	90 s
Global conflicts	859	6'598	223	2	1 s
	6'179	152'517	1'208	3	48.5 s
	22'663	706'357	3'905	4	541 s
Exoplanets	989	18'926	69	unlimited	0.8 s

Table 4: Size and performances for article neighbor subgraph requests

Page	Articles	Hyperlinks	Subcategories	Search depth	Processing time
Switzerland	1'400	144'911	24	1	4.5 s
United States	2'215	258'939	28	1	17.5 s
Charlie Chaplin	1'289	147'203	23	1	4 s
Albert Einstein	1'025	114'518	30	1	2.3 s
Computer science	684	47'067	13	1	1 s
	68'756	7'883'471	1'450	2	3'600 s

toolkit that allows researchers to retrieve this information directly from Wikipedia dumps.

Information about Wikipedia edits and editors are also available. They could be structured as a graph of articles or a graph of users with time-series of edit activity, for example.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback and useful recommendations that helped to improve this paper. V. Miz has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n^{o} 642685 MacSeNet.

REFERENCES

[1] Nicolas Aspert, Volodymyr Miz, and Benjamin Ricaud. 2019. Wikipedia dataset. https://lts2.epfl.ch/Datasets/Wikipedia/
[2] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. 2018. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 83.
[3] Francesco Bellomi and Roberto Bonato. 2005. Network analysis for Wikipedia. In *proceedings of Wikimania*.
[4] Luciana S Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. 2006. Temporal analysis of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 45–51.
[5] Ruth García-Gavilanes, Anders Mollgaard, Milena Tsvetkova, and Taha Yasseri. 2017. The memory remains: Understanding collective memory in the digital age. *Science advances* 3, 4 (2017), e1602368.
[6] Christine Klymko, David Gleich, and Tamara G Kolda. 2014. Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874* (2014).
[7] Márton Mestyán, Taha Yasseri, and János Kertész. 2013. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one* 8, 8 (2013), e71226.
[8] Volodymyr Miz, Benjamin Ricaud, Kirell Benzi, and Pierre Vandergheynst. 2019. Anomaly detection in the dynamics of web and social networks using associative memory. *The Web Conf. 2019* (2019).
[9] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. 2013. Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports* 3 (2013), 1801.
[10] Various. 2011. SNAP dataset - Wikipedia top categories. https://snap.stanford.edu/data/wiki-topcats.html
[11] Various. 2013. SNAP dataset - Wikipedia hyperlink network. https://snap.stanford.edu/data/enwiki-2013.html
[12] Robert West, Ashwin Paranjape, and Jure Leskovec. 2015. Mining missing hyperlinks from human navigation traces: A case study of Wikipedia. In *Proceedings of the 24th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1242–1252.
[13] Wikimedia. 2019. Wikipedia dumps. https://dumps.wikimedia.org/
[14] Wikimedia. 2019. Wikipedia Pageviews dumps. https://dumps.wikimedia.org/other/pageviews/
[15] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. 2012. Dynamics of conflicts in Wikipedia. *PloS one* 7, 6 (2012), e38869.
[16] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. 2017. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 555–564.
[17] Torsten Zesch and Iryna Gurevych. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*. 1–8.

Learning to Map Wikidata Entities To Predefined Topics

Preeti Bhargava*
Demandbase
San Francisco, CA
pretsbhargava@gmail.com

Nemanja Spasojevic
Youtube
San Bruno, CA
sofra@alum.mit.edu

Sarah Ellinger
Juul Labs
San Francisco, CA
sarah.ellinger@lithium.com

Adithya Rao
Magic Leap
Sunnyvale, CA
adithyar@alumni.stanford.edu

Abhinand Menon
Beeswax
New York, NY
abhinandmenon@gmail.com

Saul Fuhrmann
Lime Bikes
San Francisco, CA
frmsaul@gmail.com

Guoning Hu
Amazon
Sunnyvale, CA
guoning.hu@gmail.com

ABSTRACT

Recently much progress has been made in entity disambiguation and linking systems (EDL). Given a piece of text, EDL links words and phrases to entities in a knowledge base, where each entity defines a specific concept. Although extracted entities are informative, they are often too specific to be used directly by many applications. These applications usually require text content to be represented with a smaller set of predefined concepts or topics, belonging to a topical taxonomy, that matches their exact needs. In this study, we aim to build a system that maps Wikidata entities to such predefined topics. We explore a wide range of methods that map entities to topics, including GloVe similarity, Wikidata predicates, Wikipedia entity definitions, and entity-topic co-occurrences. These methods often predict entity-topic mappings that are reliable, i.e., have high precision, but tend to miss most of the mappings, i.e., have low recall. Therefore, we propose an ensemble system that effectively combines individual methods and yields much better performance, comparable with human annotators.

KEYWORDS

entity topic mapping; entity topic assignment; natural language processing; knowledge base; wikipedia; wikidata

ACM Reference Format:
Preeti Bhargava, Nemanja Spasojevic, Sarah Ellinger, Adithya Rao, Abhinand Menon, Saul Fuhrmann, and Guoning Hu. 2019. Learning to Map Wikidata Entities To Predefined Topics. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316749

*This work was done when all the authors were employees of Lithium Technologies | Klout

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316749

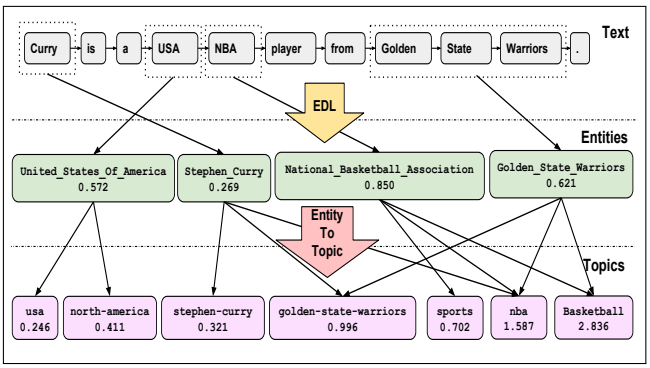


Figure 1: Topic extraction using Entity Disambiguation and Linking (EDL) together with entity-to-topic mapping.

1 INTRODUCTION

There have been many efforts to extract the rich information available in various types of user-generated text, such as webpages, blog posts, and tweets and represent it as a set of concepts which can then be used by various applications, such as content search, personalization, and user profile modeling. This can be achieved by understanding the *topics* in which users are interested or are experts [8, 16, 21, 22] by categorizing the user-generated text into a finite set of topics or categories.

Traditionally, statistical topic models such as LDA [7] have been used for topical categorization of text. These models are based on the idea that individual documents are made up of one or more topics, where each topic is a distribution over words. There have been many applications showing the power of these models on a variety of text documents (e.g. Enron emails, CiteSeer abstracts, Web pages). While LDA is a powerful tool for finding topic clusters within a document, it may miss implicit topics that are better suitable for document categorization.

Recently, tremendous advances have been made in entity disambiguation and linking (EDL) [1, 4, 9, 13, 14, 19]. Many EDL API services are now available to the public, including Google NLP[1],

[1]https://cloud.google.com/natural-language/

Watson Natural Language Understanding API[2], and Rosette Text Analytics[3]. These advanced EDL technologies and services make it practically elementary to extract a set of entities from a piece of text.

Unlike LDA-generated topics, entities are well defined concepts described in a knowledge base (KB) e.g. entities in the Wikidata KB. Modern KBs contain hundreds of thousands of entities or more. Some entities are quite broad, but more often they are very specific. When such narrow entities are extracted from text, they are somewhat informative, but they may be too specific and too many for the needs of a given application. Moreover, entities help enable a syntactic rather than a semantic understanding of the text. For example, in a search application where the query is "Golden State Warriors", documents indexed by the entity "Stephen Curry" are highly relevant, but may not be returned.

To address these challenges and meet the needs of general applications, a *topical taxonomy*, or hierarchical structure of topics, can be introduced. The primary advantage of using such a taxonomy rather than directly applying entities is to support product and business requirements such as:

(1) Limiting topics to a given knowledge domain.
(2) Imposing an editorial style or controlling the language used in describing topics (e.g., by imposing a character limit on topic names).
(3) Limiting the topic set in size so that an application user can better interact with the available topics. Topic taxonomy cardinality is orders of magnitude smaller than number of entities within the KB.
(4) Preventing unsuitable concepts from being represented as topics. These may include concepts that are:
   (a) Offensive or controversial (e.g. *Pornography*).
   (b) Either too general (e.g. *Life*) or too specific (e.g. *Australian Desert Raisin*).
   (c) Redundant with one another (e.g. *Obamacare* and *Affordable Care Act*).

For example, Klout.com[4] used a custom taxonomy[10] which was modeled around capturing topics of social media content in order to build topical user profiles [21, 22]. Another example is Google Adwords, which uses a small, human-readable taxonomy[5] to allow advertisers to target personalized ads to users based on topical interests.

Thus, to categorize text into topics, one can take advantage of mature EDL systems by *mapping* the entities extracted from the text to topics in a topical taxonomy. An example of using EDL to extract topics is shown in Figure 1. Although there have been studies that touch upon aspects of the entity-topic mapping problem, either while modeling the relationships among entities or while modeling the concepts of entities, no systematic study exists of this particular task. However, we find that an ensemble of some selected models is able to yield very good results.

Our main contributions in this paper are:

- We propose a system that maps entities in a KB (derived from Wikidata) to topics in a taxonomy. Together with EDL, our system allows one to extract the concepts that best meet specific application needs from a given text.
- We study multiple popular models that explore the relationship among entities from various perspectives, including cooccurrence, word embeddings, and Wikipedia content. We find that each of them performs reasonably well on mapping entities to topics.
- We investigate multiple approaches to combine the above models into a stacked ensemble model and obtain much better results. We find that best performance is achieved through a SVM meta-model (AUC: 0.874 and F1: 0.786) which yields results comparable to human annotators.
- We show that although our system is developed with a specific topical taxonomy, one can easily adapt it for use with other taxonomies.
- Open data - we make our label set publicly available.

2 PROBLEM SETTING

In this work, we attempt to build a system that maps entities in an entity set to topics in a topic set.

2.1 Entity set

Wikidata is the largest free and open KB, acting as a central structured data store of all Wikimedia content. Entities are the atomic building blocks of Wikidata. Information about a Wikidata entity is organized using named predicates, many of which annotate relations between entities.

We derived our entity set from Wikidata for the following reasons:

- Wikidata entities are widely used in the community, allowing our system to benefit a large audience.
- Wikidata contains more than $43M$ entities, covering the majority of concepts that people care about. These entities include people, places, locations, organizations, etc. There are entities for broad concepts, such as *Sports*, and entities for very specific concepts, such as *5th Album of The Beatles*.
- Wikidata entities come with rich annotations that can be utilized for this problem. For example:
  — Every entity is linked to a corresponding word or phrase in multiple languages.
  — Millions of predicates describe special relations among entities (see Section 5.2.1 for more details).
- There are datasets associated with Wikidata that provide useful information. In this work, we leverage Wikipedia pages (see Section 5.3.1 for more details) and DAWT [20], an extension dataset to Wikidata.

2.2 Topic set

Topics that entities are mapped to are application specific. In this study, we use a topic set from the Klout Topic Ontology (KTO) [10] which itself is a subset of Wikidata. KTO contains about 8K topics. Each topic is annotated with two types of annotations:

- **Wikidata id** - Since KTO is a subset of Wikidata, each topic is equivalent to a Wikidata entity which is referred to as the

[2]https://www.ibm.com/watson/services/natural-language-understanding/
[3]https://developer.rosette.com
[4]Klout.com has now been shut down since May 25, 2018
[5]https://support.google.com/ads/answer/2842480

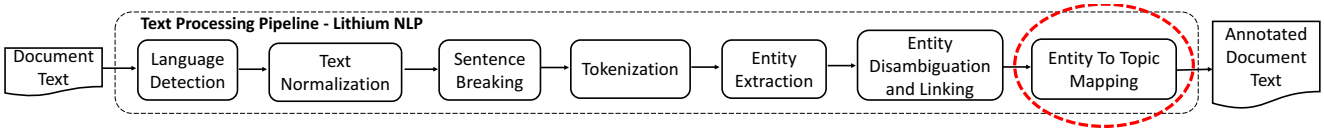


Figure 2: The Lithium NLP pipeline.

primary entity for that topic. The wikidata id for the *primary entity* is used as this annotation.

- **Parent and child topics** - Topics are organized hierarchically into multiple levels. Topics at a higher level are broader and usually link to multiple narrower topics at the next lower level. For example, topics *Entertainment*, *Pop music* and *The Beatles* have the following relationship within KTO:

  Entertainment ⇒ *Pop music* ⇒ *The Beatles.*
  This hierarchical relationship is encoded in these parent and child topic annotations.

To define our task formally, let $\mathcal{E}$ be Wikidata entity set. Let $\mathcal{T}$ be a topic set with a hierarchical structure. Let $e \in \mathcal{E}$ be an entity and $t \in \mathcal{T}$ be a topic. For any t, there is an equivalent entry in $\mathcal{E}$, which is referred to as the primary entity, e_t. Note that an entity, whether or not it is a primary entity, can map to multiple topics. To be concise, if e shall be mapped to t, we say t is relevant to e; otherwise, we say that t is irrelevant to e. Thus, the problem we attempt to solve is:

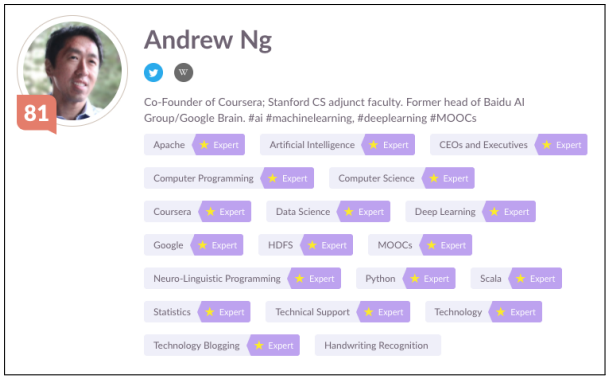
Given a pair of (e, t), determine whether or not t is relevant to e irrespective of whether t is equivalent to e

This is a binary classification problem. For practical purposes, we aim to build a regression system that not only performs classification, but also yields a quantitative measure of how relevant e is to t, which is often useful for a subsequent application stage.

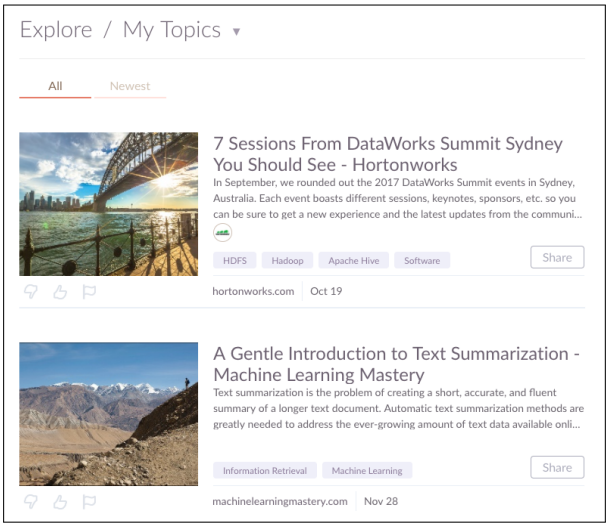
3 APPLICATIONS

The described system is a part of the Lithium NLP [6] [5] pipeline - a resource-constrained, high-throughput and language-agnostic system for information extraction from noisy user generated text such as that available on social media. Figure 2 shows a partial view of the Lithum NLP pipeline, where the sub-module 'Entity To Topic Mapping' is the final stage. Lithium NLP is capable of extracting a rich set of information including entities, topics, hashtags and sentiment. Lithium NLP currently supports multiple languages including Arabic, English, French, German, Italian and Spanish. It supports large scale data from several social media platforms such as Twitter, Facebook, Linkedin, etc. by processing about $500M$ new social media messages, and $0.5M$ socially relevant URLs shared daily. Since it employs statistical NLP techniques, it uses the large scale of the data to help overcome the noisiness.

In the Lithium NLP pipeline, entity-to-topic mapping is used to convert a set of Wikidata entities to a set of topics that are most relevant to a piece of text. These topics are stored as annotations, which are consumed by multiple Lithium products for various tasks, such as indexing content and building user profiles. One notable application was Klout[7] - a consumer platform which integrated users' data from multiple social networks in order to measure their online social influence via the *Klout Score*[8] [18]. On Klout, these topic mappings were used to model users' topics of interest [22] and expertise [21] in order to recommend personalized content to the users. Figure 3a shows a user's topical profile on Klout, and Figure 3b shows content recommendations derived from those topics. In addition, these topical profiles were included in Twitter's PowerTrack APIs[9].



(a) Andrew Ng's inferred expertise topics.



(b) Andrew Ng's content feed, personalized to his topics.

Figure 3: Topic mapping enabled applications.

[6] http://nlp.app.lithium.com/ui

[7] https://klout.com

[8] https://klout.com/corp/score

[9] http://support.gnip.com/enrichments/klout.html

Lithium NLP currently enables text and user annotations within Lithium's social media management tools[10], and is used to analyze $20 + M$ new daily engagements across Lithium's 400+ communities[11].

4 DATA SETS

As described in Section 2, we use an entity set derived from Wikidata as our entity set and KTO as our topic set. Wikidata entities were limited to the $1M$ most important ones, where importance measure is described in Bhattacharyya and Spasojevic [6]. To train our ensemble, we collected a dataset of $26.6K$ entity-topic pairs labeled as 'Relevant' if the topic is relevant to the entity or 'Irrelevant' otherwise. We had 3 data sources - a control dataset which we used to control quality of operator evaluations and two labeled datasets that we used for training, validation, and testing.

4.1 Control Set

To control the quality of labeling done by annotators, a small set of 100 labeled $(entity, topic)$ pairs was generated by an in-house expert panel. 10% of each batch of tasks prepared for human annotators included data taken from this control set; based on how accurate that human's evaluations were for control-set pairs, we were able to estimate the quality of the batch as a whole.

4.2 Amazon Mechanical Turk

We collected about $8.6K$ annotations via Amazon Mechanical Turk (AMT). Workers were shown an entity and its description as well as a topic and its description; they were then asked to identify whether the entity hierarchically mapped to the topic, or was unrelated. The entities were picked from the top $100K$ entities of our KB. For each entity-topic pair within Wikidata and its predicate, we sampled pairs so that predicate distribution was balanced. This guaranteed that a diverse set of relationships was present in our data set.

Each AMT task had about 100 such entity-topic pairs and was given to one unique worker, who was compensated $3. Before doing large-scale data collection, we calibrated the workers on accuracy and time via a pilot set using 3 workers. The workers exhibited an accuracy of 80% when compared to the control set labels. Hierarchical relationships between entities and topics is a subjective measure and as such there may not be total agreement among the workers. Hence, we also computed the consensus of workers on the pilot test data and found that they showed a consensus of 81%.

4.3 In-House Labeling

We also collected labeled data from 5 team members labeling about $18K$ entity topic pairs. In this case we sampled data from the top $1M$ entities where the probability of selecting an entity was inversely proportional to its rank. This ensured that we have higher representation of top-ranked entities, which are more important to our application. As with Mechanical Turk, we did a pilot test to see how accurately team members labeled the data. Our in-house experts achieved an accuracy of 89% on the pilot test and a consensus of 82%.

[10] https://www.lithium.com/products/social-media-management/
[11] https://www.lithium.com/products/online-communities/

5 METHODOLOGY

Figure 4 shows a high level overview of our entity-topic mapping pipeline. The pipeline takes an (entity, topic) pair and yields a measure of how relevant the topic is to the entity. To build this system, we first considered a wide range of models that approached this problem from different aspects. Some of them generated statistical metrics on mapping an entity to a topic. Others were models that had been applied to similar problems erstwhile.

We then combined these models using a stacked ensemble as this helps overcome the inherent biases and lack of coverage (manifested as low recall) of the individual models. To make the system run efficiently in production, we removed models that were either too computationally intensive or had insignificant contributions to the performance[12]. The final ensemble combines 8 models, which are described below.

As a running example through this section, we will use the entity *Q1414593 - Patient Protection and Affordable Care Act*. Table 1 shows the mapping scores generated by the 8 models between this entity and topics that are most relevant to it.

5.1 Word Embedding Based Methods

5.1.1 GloVe Model. Pennington et al. [15] introduced Global vectors for word representation (GloVe) - a technique that generates word embeddings, i.e., vectors of real numbers, for representing the linguistic contexts of words. These embeddings allow the derivation of quantitative distance metrics, such as cosine similarity, to describe semantic "distance" between words.

As part of this model, we calculate the cosine similarity of embeddings for each (entity, topic) pair by using the primary entity for the topic. Thus, given a topic t, let e_t be the corresponding primary entity, we then have a score S measuring the similarity between t and an entity e as

$$S(t, e) = S(e_t, e). \tag{1}$$

5.1.2 GloVe Parents model. Although an entity and a relevant topic tend to have high word embedding based similarity, an entity and an irrelevant topic may often have high similarity too. Take the example of the entity *San Francisco* and the topics *Los Angeles* and *California*. Similarity between *San Francisco* and *Los Angeles* is 0.66, whereas that between *San Francisco* and *California* is 0.55. While this similarity metric indicates *Los Angeles* as a better topic for *San Francisco*, because both *Los Angeles* and *San Francisco* are cities in *California*, their relationship is not hierarchical, which is why *California* would be the only acceptable topic.

Such hierarchical relationships are not well represented in GloVe vectors, which rather capture peer to peer similarity. However, these hierarchical relationships are implicitly represented in our topic set $\mathcal{T}$. In particular, our topic set is structured as a directed acyclic graph, where the edges between topics represent hierarchical relationships. Topics in a higher level are broader and usually link to multiple narrower child topics in the next lower level. For example, both *Los Angeles* and *San Francisco* are child topics of *California*.

Therefore, we combine embeddings with the hierarchy of our topic set $\mathcal{T}$ to build a new measure. We first pick the most relevant

[12] Due to lack of space, we are unable to describe the models that we removed.

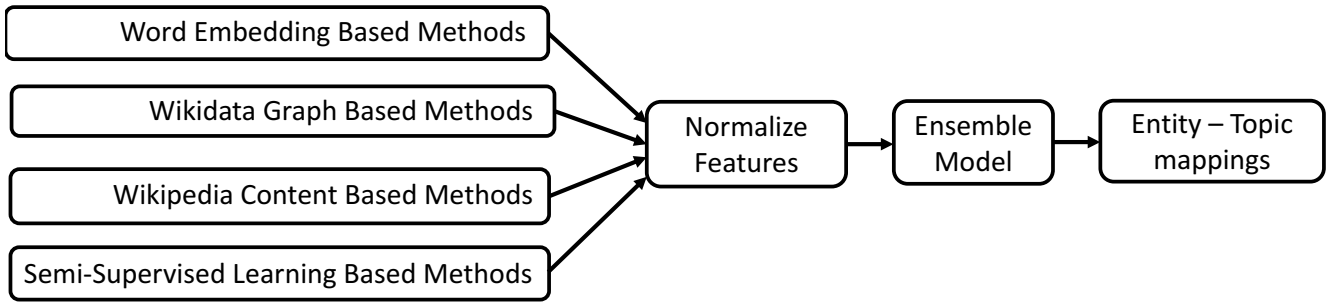


Figure 4: High level overview of the entity-topic mapping pipeline.

Table 1: Entity-topic mappings generated for the entity "Patient Protection and Affordable Care Act"

Model	Health Care	Health Insurance	U.S. Presidents	Politics	Affordable Care Act
GloVe	0	0	0.462	0.535	1.0
GloVe Parents	0.264	0.755	0.277	0.298	0
Wikidata Hierarchical Predicates	0	0	0	0	1.0
Wikidata Hierarchical Location Predicates	0	0	0	0	0
Wiki Pages Content	0.866	0	0.701	0.338	1.0
Wiki Pages Content Parents	0.132	0.868	0.229	0.736	0
Frequency Adjusted Co-occurrence	0	0	0	0.447	1.0
Topic Normalized Co-occurrence	0	0	0	0	1.0
Combined Ensemble (SVM)	0.905	0.904	0.896	0.896	1.0

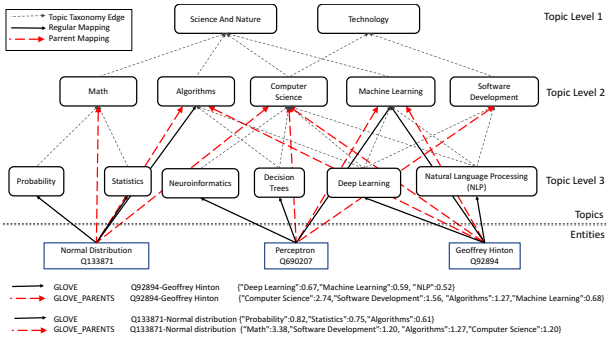


Figure 5: Higher-level topics inferred using the topic hierarchy.

topics via above similarities. Let V be the set of K topics that is most relevant to entity e. That is,

$$S(t_i, e) \geq S(t_j, e) \; \forall \; t_i \in V, t_j \notin V \tag{2}$$

We then calculate a combined similarity score of a parent topic t_p from the similarity scores of its child topics. Let C be the set containing all child topics of t_p. The combined similarity score between t_p and e is:

$$S(t_p, e) = \sum_{t_i \in C \cap V} S(t_i, e) \tag{3}$$

Parameter K is obtained through cross-validation and best performance is obtained when $K = 10$.

As an example, Figure 5 demonstrates how we leverage the topic hierarchy to infer the higher level topics for an entity. For entity *Q133871-Normal distribution*, the GloVe word vector model generates a ranked list of topics as: { *Probability* : 0.82, *Statistics* : 0.75, *Algorithms* : 0.61 .... }. Topic *Probability* and *Statistics* have parent topic *Math*. In the GloVe word vectors parents model, we aggregate the strengths from the children topics to infer {*Math* : 3.38, *Software Development* : 1.20, *Algorithms* : 1.27, *Computer Science* : 1.20....} for the entity *Q133871-Normal distribution*.

5.2 Wikidata Graph Based Methods

5.2.1 Wikidata Hierarchical Predicates Model. In this model, we leverage the structure of the Wikidata graph to infer hierarchical relationships between entities and topics. Although there are thousands of predicates defined in Wikidata, most (entity, topic) pairs have zero or one predicate. As a result, we only observe 146 unique predicates between entities and topics in our data set. In addition, we notice that location-related topics are overrepresented in this set. Hence, we consider only non-loccation predicates.

First, for each entity topic pair (e, t), let us define a predicate vector as $P(e, t) = \{r_i(e, t)\}, i \in [1, 146]$. $r_i(e, t)$ is 1 if the corresponding predicate connects e and t and 0 otherwise. We then apply a logistic regression model to estimate the probability of mapping e to t from predicate vector:

$$H(e, t) = \frac{1}{1 + e^{-\theta^t P(e,t)+b}} \tag{4}$$

Weight vector θ and bias b are estimated using the training dataset discussed in Section 8. As expected, individual weights assigned to predicates are fairly interpretable such that predicates with high score tend to correspond to hierarchical relations, such as "Instance Of" (P17), "Subclass of" (P279) etc.

5.2.2 WikiData Hierarchical Location Predicates Model. In this model we consider only the location predicates which are excluded from the previous model. This new location-based model is a Logistic Regression model.

5.3 Wikipedia Content Based Methods

5.3.1 Wikipedia Pages Content Model. The Wikipedia page of an entity usually gives a definition that contains related topics in the first paragraph. Therefore, we derive a measure that links an entity to topics based on its Wikipedia definitions in multiple languages.

Given an entity e, we first extract primary entities from the first paragraph of the corresponding Wikipedia page P_l for a given language l, via the EDL algorithm described in Bhargava et al. [4]. As an example, below is the first paragraph of the entity *machine_learning* in the English Wikipedia with extracted primary entities in bold:

*Machine learning is a field of **computer science** that gives **computers** the ability to learn without being explicitly programmed.*

Thus, for a topic t with primary entity e_t, the similarity score (per language) is defined as

$$S_l(t,e) = \begin{cases} \frac{1}{d(e_t)} & if\ e_t \in P_l \\ 0 & else \end{cases} \tag{5}$$

where $d(e_t)$ is the distance from the n-gram representing e_t to the beginning of the paragraph. The derived measure is an aggregation of the above score across all languages:

$$S(t,e) = \sum_l S_l(t,e) \tag{6}$$

Leveraging this across multiple languages helps boost most relevant topics for a given page. Note that since Wikipedia pages often contain location topics (eg. place of birth, location of headquarters, etc), which may often be irrelevant, we remove all the location topics in this measure.

5.3.2 Wikipedia Pages Content Parents Model. Similar to the GloVe word vector parents model described in Section 5.1.2, we combine the output of the Wikipedia pages content model with the inherent hierarchy of our topic set $\mathcal{T}$. Given an entity e, let V_d be the set of K topics that is most relevant to e. That is,

$$S(t_i,e) \geq S(t_j,e)\ \forall\ t_i \in V_d, t_j \notin V_d \tag{7}$$

We calculate a combined score of a parent topic t_p from the wiki-content scores of its child topics. Again, let C be the child topic set of topic t_p. The combined similarity score between t_p and e is:

$$S(t_p,e) = \sum_{t_i \in C \cap V} S(t_i,e) \tag{8}$$

5.4 Semi-Supervised Learning Based Methods

Our final model for entity to topic mapping utilizes co-occurrence frequencies for entities obtained from our DAWT [20] dataset. We capture co-occurrence frequencies among entities by counting all the entities that simultaneously appear within a sliding window of 50 tokens. Moreover, this data is accumulated across all languages and is language-independent in order to better capture relations and create a smaller memory footprint when supporting additional languages. Also, for each entity, we consider only the top 30 co-occurring entities which have at least 10 or more co-occurrences across all supported languages.

5.4.1 Frequency Adjusted Co-occurrence. Co-occurrence count has its own shortcomings as a measure of quantifying relationship between entities, because some entities occur more frequently in the dataset than others and hence will co-occur more with other entities. Therefore, we adjust the co-occurrence counts with entity frequencies. Let e denote an entity, $C(e)$ denote the occurrence count of e, N denote the total number of entities, and $C(e_i,e_j)$ denote the co-occurrences count of entity e_i and e_j. Then, the frequency adjusted co-occurrence of e_i and e_j i.e. $C_f(e_i,e_j)$ is calculated as:

$$C_f(e_i,e_j) = C(e_i,e_j) \cdot \log\left(\frac{N}{C(e_j)}\right) \tag{9}$$

Thus, given a topic t and e_t as its equivalent primary entity, we have a measure of linking an entity e to a topic t via $C_f(e,e_t)$.

5.4.2 Topic Normalized Co-occurrence. Some topics happen more frequently than other topics. As a result, frequency adjusted co-occurrences tend to favor topics that occur often in our corpus and knowledge base. To address this problem, we introduce a measure that further normalizes frequency adjusted co-occurrence for each topic. In particular, $C_t(t,e)$, normalized co-occurrence of topic t and entity e, is calculated as:

$$C_t(t,e) = \frac{C_f(t,e)}{\sum_i C_f(t,e_i)} \tag{10}$$

In this model, for each primary entity, we use the *co-occurring entity* with the strongest weight for the frequency adjusted co-occurrence. We then aggregate the topics for each *co-occurring entity* by mapping the primary entity to its equivalent topic. These give us the most relevant topics for the *co-occurring entity* with the normalized frequency adjusted co-occurrence score as the strength of the entity topic mapping.

5.5 Stacked Ensemble Model

To frame this as a machine learning problem, we define a feature as a numerical score associated with how strongly an entity should map to a topic in a hierarchical manner. Each of the models described above defines a separate feature. Our goal is to combine these features in a stacked ensemble to give the least error or loss in terms of entity to topic mappings.

We wish to compute a score for each entity-topic pair, trying to estimate the relevance of the topic to the entity. We define a feature vector $\mathcal{F}(t,e)$ for a entity e and a topic t as:

$$\mathcal{F}(t,e) = [f_1(t,e), f_2(t,e), ..., f_m(t,e)] \tag{11}$$

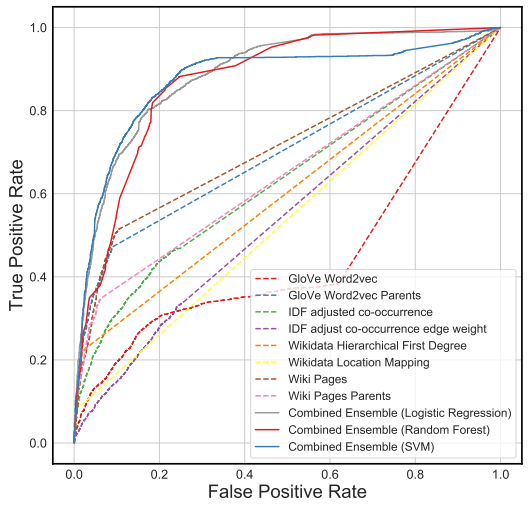


Figure 6: ROC curve on test set.

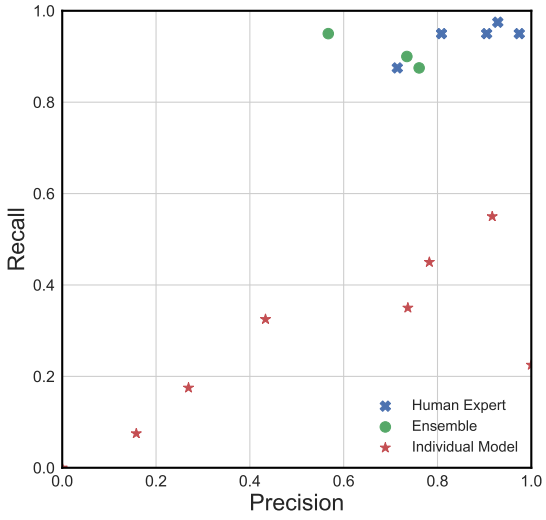


Figure 7: Precision and recall on control set.

where $f_k(t,e)$ is the feature value associated with a specific model which were introduced earlier. The normalized feature values are denoted by $\hat{f}_k(t,e)$ and the normalized feature vector is represented as: $\hat{\mathcal{F}}(t,e) = [\hat{f}_1(t,e), \hat{f}_2(t,e), ..., \hat{f_m}(t,e)]$. Normalization was calculated as:

$$\hat{f}_k(t,e) = \frac{f_k(t,e)}{\max_{t_j \in \mathcal{T}} f_k(t_j,e)} \tag{12}$$

All training has been performed on normalized feature vectors.

We then treat this task as a binary classification problem and apply the following methods: logistic regression, random forest, and support vector machine (SVM). The results are summarized in the Section 6.

6 EVALUATION AND RESULTS

In Table 2 we show the performance of individual models as well as ensemble models. The evaluations were performed on the reserved test set, which represented 20% of the labeled dataset not including the *control set*. We can see that each individual model performs reasonably well on this task. The best performing models, *Wikipedia Pages* and *GloVe Parents*, had an AUC of 0.721 and 0.701 respectively, while the worst performing model was *GloVe* with an AUC of 0.449. The poor performance of GloVe can be explained by the fact that GloVe embeddings capture similarity well, but not hierarchy.

At the bottom of Table 2 we show the performance of the ensemble models. All of the ensemble models had roughly similar performance, and significantly outperformed individual models. The SVM ensemble performed the best with a F1 of 0.786 and an AUC of 0.874. The ROC curves for individual and ensemble models are shown in Figure 6. We notice also that the majority of performance deterioration for individual models is caused by low recall. All of the models have recall less than 0.325, while the best ensemble achieved recall of 0.752. This means that we have chosen a diverse set of models, complementary to each other, which when combined result in high performance.

In Figure 7 we plot precision and recall for individual models, ensemble models, and expert human annotators evaluated on the *control* introduced in Section 4.1. We can see that the ensemble models were comparable to human annotator performance, while individual models were less successful.

7 LESSONS LEARNED

Some of the challenges we faced when implementing our entity to topic mapping system were:

- **Ambiguity of Relationships.** Topic mapping is very subjective; the perceived strength of the relationship between an entity and a topic can vary from human to human, depending on their domain knowledge and the intended application. For example, should the music group *Deep Purple* map to *Progressive Rock*, although their musical genre shifted over time toward *Heavy Metal*? These examples demonstrate the subjectivity and ambiguity of the problem.
  In addition, we found that Mechanical Turk annotators reached only 81% consensus, with in-house annotators reaching 89% when measured on the *control set*, thus, demonstrating the ambiguity.
- **Similarity vs. Hierarchy.** Many models successfully capture similarity (peer-to-peer relations) between an entity and a topic; however, parent-child relations are much harder to capture. To minimize false positives, we introduced the *Parent* models that take a base model and roll up candidate topics to their parents within the topic taxonomy.
- **Knowledge Base Biases.** Many KBs are rich with factual data; however, some entity subsets may be overrepresented for the purposes of an application. For example, we notice

Table 2: Performance metrics for the different models.

Model	F1	AUC	Recall	Precision	Accuracy
GloVe	0.333	0.449	0.232	0.590	0.606
GloVe Parents	0.429	0.701	0.282	0.895	0.681
Wikidata Hierarchical Predicates	0.286	0.599	0.170	0.917	0.641
Wikidata Hierarchical Location Predicates	0.142	0.536	0.077	0.912	0.605
Wiki Pages Content	0.456	0.721	0.312	0.852	0.685
Wiki Pages Content Parents	0.328	0.641	0.203	0.850	0.647
Frequency Adjusted Co-occurrence	0.440	0.635	0.325	0.680	0.649
Topic Normalized Co-occurrence	0.224	0.553	0.142	0.530	0.583
Combined Ensemble (Logistic Regression)	0.760	0.887	0.695	0.839	0.806
Combined Ensemble (Random Forest)	0.757	0.871	0.735	0.780	0.791
Combined Ensemble (SVM)	0.786	0.874	0.752	0.824	0.819

that in Wikipedia and Wikidata entities locations are frequently described in even non-geographic entities. Although still informative for our task, the frequency of these relations had a negative impact on final results. To address this problem we must either filter location-based topics or split a single model into two disjoint models, where one represents mappings purely to location-related topics and other represents mappings to rest of taxonomy.

- **Taxonomy Constraints.** In our problem statement we said that topic taxonomy is subset of KB entities, and based on this principle, we heavily rely on a 1:1 mapping of a topic to its primary entity. However, depending on the taxonomy used, there may be cases where a topic does not have a representative entity, in which case a workaround would have to be devised.

8 OPEN DATASET

The dataset used to run evaluations and build models has been opened at **https://github.com/klout/opendata**. It includes $26.6K$ triplets $(entity, topic, label)$ where for a given pair $(entity, topic)$ $label$ indicates if $entity$ should map to given $topic$ or not. A detailed explanation of how data has been sampled and how labels have been generated can be found in Section 4. Each entity is represented with a Wikidata[13] id and each topic with a Klout Topic Ontology id [10]. For convenience, the dataset also includes a display name.

9 RELATED WORK

Statistical topic modeling such as LDA [7] is based on the premise that individual documents are made up of one or more topics and each topic is a distribution over words. Newman et al. [12] adapt LDA to model associations between entities and topics that occur in a document. Kim et al. [11] propose a topic model for analyzing a collection of documents with given entities and model the correlation between entity-term and topic-term distributions.

However, we are primarily interested in finding mapping a set of entities to a set of topics regardless of the documents that the entities appear in. We can think of these mappings as more global mappings that are not local to the documents and do not depend on the document. This problem has not received significant attention in the industry. Balog [2] presents a model of entity-topic mapping in order to identify topical experts but they focus on only 2 types of entities - people and moods. Balog et al. [3] extend this work to identify entity topic mappings in online news articles. Raghavan et al. [17] explored entity-entity modeling and relationship description but mostly focused on entities that co-occur together in a text window. This is similar to the semi-supervised learning models in our ensemble model.

Our work differs from these previous as we attempt to build a generic system that maps Wikidata entities to any predefined topic taxonomy.

[13] https://www.wikidata.org

10 CONCLUSION AND FUTURE WORK

In this paper, we addressed the problem of mapping Wikidata entities to predefined topics. We built an ensemble model that leverages GloVe word vector similarity, Wikidata predicates, Wikipedia entity definitions, and entity-topic co-occurrences for mapping entities to topics. Our system obtains performance comparable to human annotators and has been integrated as part of the Lithium NLP pipeline, serving multiple applications running in production. We continually collect production data and customer feedback to further improve the system. In future, we plan on incorporating multiple KBs to handle Knowledge Base bias. In addition, as this research was done using a single topic set, we would like to test against other domain-specific taxonomies, and extend our system to handle cases where there is no 1:1 topic to entity mappings.

REFERENCES

[1] Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. 2014. Automatic Creation of Arabic Named Entity Annotated Corpus Using Wikipedia. Association for Computational Linguistics.

[2] Krisztian Balog. 2007. People search in the enterprise. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 916–916.

[3] Krisztian Balog, Maarten de Rijke, Raymond Franz, Hendrike Peetz, Bart Brinkman, Ivan Johgi, and Max Hirschel. 2009. Sahara: Discovering entity-topic associations in online news. In *Proceedings of ISWC*, Vol. 9.

[4] Preeti Bhargava, Nemanja Spasojevic, and Guoning Hu. 2017. High-Throughput and Language-Agnostic Entity Disambiguation and Linking on User Generated Data. In *Proceedings of WWW 2017 workshop on Linked Data on the Web*.

[5] Preeti Bhargava, Nemanja Spasojevic, and Guoning Hu. 2017. Lithium NLP: A System for Rich Information Extraction from Noisy User Generated Text on Social Media. In *Proc. of the 3rd Workshop on Noisy User-generated Text*. 131–139.

[6] Prantik Bhattacharyya and Nemanja Spasojevic. 2017. Global Entity Ranking Across Multiple Languages. In *Companion Proceedings of the WWW*. 761 – 762.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[8] Christopher S Campbell, Paul P Maglio, Alex Cozzi, and Byron Dom. 2003. Expertise identification using email communications. In *Proc. of ACM Conference on Information and Knowledge Management (CIKM)*.

[9] Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data.. In *EMNLP-CoNLL*, Vol. 7. 708–716.

[10] Sarah Ellinger, Prantik Bhattacharyya, Preeti Bhargava, and Nemanja Spasojevic. 2017. Klout topics for modeling interests and expertise of users across social networks. *arXiv preprint arXiv:1710.09824* (2017).

[11] Hyungsul Kim, Yizhou Sun, Julia Hockenmaier, and Jiawei Han. 2012. Etm: Entity topic models for mining documents associated with entities. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 349–358.

[12] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. 2006. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 680–686.

[13] Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. AIDA-light: High-Throughput Named-Entity Disambiguation.. In *LDOW'14*.

[14] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194 (2012), 151–175. https://doi.org/10.1016/j.artint.2012.03.006

[15] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation.. In *Proceedings of EMNLP*. 1532 – 1543.

[16] Ana-Maria Popescu, Krishna Yeswanth Kamath, and James Caverlee. 2013. Mining Potential Domain Expertise in Pinterest.. In *UMAP Workshops*.

[17] Hema Raghavan, James Allan, and Andrew McCallum. 2004. An exploration of entity models, collective classification and relation description. In *KDD Workshop on Link Analysis and Group Detection*. 1–10.

[18] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor Dsouza. 2015. Klout score: Measuring influence across multiple social networks. In *IEEE Intl. Conf. on Big Data*.

[19] Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*. Springer, 93–115.

[20] Nemanja Spasojevic, Preeti Bhargava, and Guoning Hu. 2017. DAWT: Densely Annotated Wikipedia Texts across multiple languages. In *Companion Proceedings of WWW*. 1655 – 1662.

[21] Nemanja Spasojevic, Prantik Bhattacharyya, and Adithya Rao. 2016. Mining half a billion topical experts across multiple social networks. *Social Network Analysis and Mining* 6, 1 (2016), 1–14.

[22] Nemanja Spasojevic, Jinyun Yan, Adithya Rao, and Prantik Bhattacharyya. 2014. LASTA: Large Scale Topic Assignment on Multiple Social Networks. In *Proc. of ACM KDD*. 1809 –1818.

Characterizing and comparing Portuguese and English Wikipedia medicine-related articles

Gil Domingues
INESC TEC
Faculty of Engineering of the University of Porto
Porto, Portugal
gil.domingues@fe.up.pt

Carla Teixeira Lopes
INESC TEC
Faculty of Engineering of the University of Porto
Porto, Portugal
ctl@fe.up.pt

ABSTRACT

Wikipedia is the largest on-line collaborative encyclopedia, containing information from a plethora of fields, including medicine. It has been shown that Wikipedia is one of the top visited sites by readers looking for information on this topic. The large reliance on Wikipedia for this type of information drives research towards the analysis of the quality of its articles. In this work, we evaluate and compare the quality of medicine-related articles in the English and Portuguese Wikipedia. For that we use metrics such as authority, completeness, complexity, informativeness, consistency, currency and volatility, and domain-specific measurements, in order to evaluate and compare the quality of medicine related articles in the English and Portuguese Wikipedia. We were able to conclude that the English articles score better across most metrics than the Portuguese articles.

CCS CONCEPTS

• **Information systems** → **Wikis**; • **Applied computing** → **Consumer health**; *Health informatics.*

KEYWORDS

Information Quality, Wikipedia, Medicine-related Content, Cross-Language Information Retrieval.

ACM Reference Format:
Gil Domingues and Carla Teixeira Lopes. 2019. Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3316758

1 INTRODUCTION

Wikipedia is a collaborative, open-source, on-line encyclopedia where users from around the world can freely contribute with information on a wide range of topics, in several languages. As of 2018, it is ranked as the fifth most popular site in the world in the Alexa ranking [2].

This editorial openness characteristic of Wikipedia has its advantages and its drawbacks. While anyone is allowed to edit articles, effectively sharing their knowledge with the rest of the world, this freedom also opens doors to vandalism and incorrect, unfounded information. However, it has been shown that Wikipedia manages to be comparable in quality with other print encyclopedias [4], and that the speed at which its articles can be updated enables the quick neutralization of vandalism [6].

Although medicine articles may not represent the largest portion of Wikipedia articles (over 33,000 out of a total of over 5,700,000 articles in the English Wikipedia) [1], the relevance of these articles cannot be understated. Wikipedia's medicine articles rank highly in the majority of the popular search engines, managing to surpass other resources, such as MedlinePlus and NHS Direct Online [8], which are domain-specific.

Given the importance of the Wikipedia articles of the medical area to the readers, there is a significant lack of analysis of its quality, more prevalent in languages other than English [5].

The goal of this paper is to use previously suggested information quality metrics applied to the Wikipedia domain, as well as other measurements we considered relevant, to analyze and compare a range of characteristics in Wikipedia's medicine articles, for the English and the Portuguese languages.

2 RELATED WORK

As the largest collaborative encyclopedia on-line, Wikipedia has, over time, attracted much attention in the research community, namely to the quality of its contents. According to Blumenstock [3], the quality of the articles is correlated with their word count. By selecting a sample of random articles and another sample of *featured* articles - articles selected by the editors as articles of great quality - , the word count of the articles of each group was determined and was demonstrated to be a good indicator of article quality, as articles with a length above a certain threshold were more likely to be correctly classified as *featured* articles.

On the other hand, Zeng et al. [13] rely on information present in the revision history of an article to predict its quality. It builds a revision history-based trust model using a dynamic Bayesian network.

Stvilia et al. [12] use a different approach to analyze the quality of Wikipedia pages. It adapts a set of previously defined metrics for information quality on the web and adapts them to the context of Wikipedia. These metrics - *Authority, Completeness, Complexity, Informativeness, Consistency, Currency and Volatility* - represent different facets of each article, thus providing a more intricate representation of the articles' quality.

As mentioned earlier, research has shown that Wikipedia is one of the most accessed on-line resources by users looking for medical

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316758

information, ranking highly in the most popular search engines [8].

Heilman et al. [5] analyze four components of medicine articles in Wikipedia: the amount of medical content in Wikipedia, the citations supporting this content, the user traffic for articles on this topic and characteristics of the users who contribute to these articles. This work did not focus on article quality itself, rather on the state of the medical content on Wikipedia across languages, its quantity, readership and the characteristics of its contributors.

Some other works have analyzed the quality of Wikipedia articles in more specific fields of medicine, such as Thomas et al. [11], which evaluated Wikipedia as an educational tool for patients in the field of nephrology, and Kräenbring et al. [7], which studied the quality of drug related information in Wikipedia. Both these works reached similar conclusions that Wikipedia is a fairly reliable and comprehensive source of information in their specific fields.

3 METHODOLOGY

In order to assess the quality of the medicine related articles in both the Portuguese and the English Wikipedia's, an approach consisting of two major steps was followed.

Firstly, we selected a dataset of Portuguese and English medicine related articles and collected the contents of these articles as well as some metadata related to their revision history, editors, content and structure.

Secondly, we chose a set of metrics and measurements that could, on the one hand, help determine the quality of the articles in the dataset and, on the other hand, describe how some of the medical information was structured in these articles.

3.1 Data Collection

The data collection process will be outlined in this section. Figure 1 provides a simplified diagram of the process, from the collection of the dataset and obtaining the article data and metadata from the Wikipedia API, to the process of calculating the metrics and measurements.

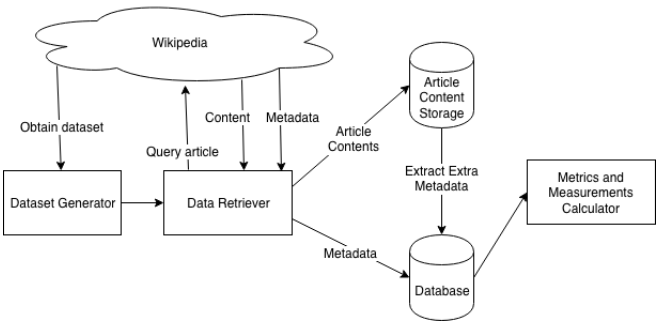


Figure 1: Outline of the process of data collection, from the creation of the dataset to the calculation of the metrics and measurements.

To obtain the English dataset we have used the list of top viewed 5,000 articles of the English Wikipedia compiled by user *West.andrew.g*[1].

[1]https://en.wikipedia.org/wiki/User:West.andrew.g/Popular_medical_pages

The Portuguese dataset was obtained by following the Portuguese language link in each of the articles of the original English dataset. As some of the pages did not have a Portuguese version, this latter dataset is smaller than the former, consisting of 3,496 articles.

The article data and metadata was obtained through the use of the MediaWiki API. Through this API, not only the current state of the article's contents was obtained, but also metadata of the article, information on the revision history, language links, wiki internal links and external links. All the resulting data was stored in a relational database. Figure 2 shows the structure of the database where the data was stored.

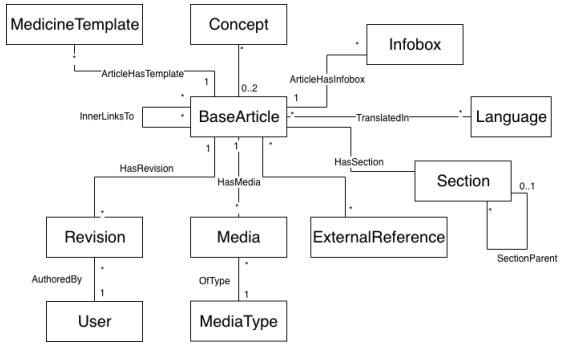


Figure 2: Structure of the database where the data obtained from Wikipedia was stored.

The remaining data necessary to calculate the metrics was obtained from processing the article's Wikipedia markup. For instance, the media files metadata was gathered through the article's markup as the API's method to obtain the images returns images which are not relevant for the content of the article, such as the lock image in protected pages. When gathering this data from the markup text those images can be filtered out. Templates and infoboxes were also extracted from the markup text, as well as the citations. This information was also added to the database.

To obtain measurements related to text analysis, such as article length, the Flesch reading ease and *InfoNoise* - a function of the number of tokens after stemming and stopping and the article length -, all the markup was removed from the article's content and a plain text file with said content was generated as a result. These measurements were then collected from this plain text file.

In order to only gather data on medicine related infoboxes and templates, two lists of relevant templates were generated. One was composed of all the templates belonging to the category *Medicine Templates* or to its subcategories, while the other was composed of all the infoboxes belonging to the category *Medicine infobox templates* or to its subcategories. Later, when analyzing every article, every template which appeared on the article and on the template list, as well as every infobox which appeared on the article and on the infobox list, were inserted into the database. To obtain the infoboxes and templates in Portuguese, the language links were followed for each entry.

A similar approach was followed to determine which users were administrators. The list of administrators from the English

Wikipedia[2] and the Portuguese Wikipedia[3] were obtained from the pages listing these users. As mentioned before in the metric formulas, this list was required to calculate the consistency metric.

3.2 Metrics

In order to compare the quality between articles, metrics on information quality had to be selected. We resorted to IQ - information quality - metrics previously defined in literature. We used the metrics as defined by Stvilia et al. [12], as these metrics accurately represented distinct components of the articles, being more comprehensive than using a single measurement to evaluate information quality, such as article length.

For the purpose of convenience, the following list describes how each metric is calculated, as described by Stvilia et al [12]. [12]:

- **Authority** = 0.2 * Num. Unique Editors + 0.2 * Total Num. Edits + 0.1 * Connectivity + 0.3 * Num. of Reverts + 0.2 * Num. External Links + 0.1 * Num. Registered User Edits + 0.2 * Num. Anonymous User Edits.
- **Completeness** = 0.4 * Num. Internal Broken Links + 0.4 * Num. Internal Links + 0.2 * Article Length
- **Complexity** = 0.5 * Flesch reading ease - 0.5 * Kincaid grade level.
- **Informativeness** = 0.6 * InfoNoise - 0.6 * Diversity + 0.3 * Num. Images.
- **Consistency** = 0.6 * Admin. Edit Share + 0.5 * Age.
- **Currency** = Currency
- **Volatility** = Median Revert Time

These metrics are calculated from a set of measurements obtained from each article. Some of these measurements are simple to obtain, while others have to be computed and are not as straightforward.

When calculating the connectivity measurement, which consists of the number of articles connected to an article via common editors, we faced an implementation obstacle. As we only collected the revisions' history from the articles included in the datasets, the calculation of connectivity was limited to the articles in the dataset. Given the large dimension of the datasets, we assume this approximation will not deteriorate the metric. From an optimistic perspective, it will strengthen the concept of authority as the connected articles will necessarily be related to the topic of medicine.

Analyzing texts in two languages may cause issues with the *Complexity* IQ metric, as the process of calculating the Flesch reading ease and the Flesch-Kincaid grade level are not applicable to Portuguese as they are to English. There is a counterpart to the Flesch reading ease in Portuguese, consisting of adding 42 to the result of the text's Flesch reading ease [10], which was used. However, as far as we are aware, there is no equivalent in Portuguese to the Kincaid grade level. To maintain a *Complexity* metric, we consider its value equal to the article's Flesch reading ease value. While this modification may seem arbitrary, this formula for *Complexity* was present in an earlier version of Stvilia et al. [12] and as such we consider it a reasonable approximation.

The metrics were defined in order to be generic enough to be applicable to an article of any topic in Wikipedia. However, in this paper we intend to analyze articles from the medicine domain, which motivated the use of a complementary set of measurements, specifically relevant to this domain, which we used to compare the specificity and detail of the medical content of the articles.

One of the elements used to structure information in Wikipedia are templates. Templates are embedded pages that allow the repetition of information. These are examples of very useful components, which can improve the manner in which the readers obtain information and have not been considered in the metrics previously described. Figure 3 shows an example of a template, taken from the influenza page on the English Wikipedia. This particular template is used to store links to external medical resources and classifications.

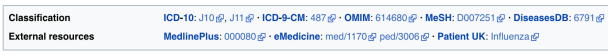


Figure 3: Template present in the English wikipedia page for influenza.

One of the most recurring types of template are infoboxes, which provide the user with a way to read critical information in a compact form. These structures are usually positioned on the top of the page, containing a summary of important information and sometimes figures. In figure 4 we can see an example of an infobox, also taken from the influenza page on the English Wikipedia. This infobox contains critical information such as symptoms, causes, usual onset, among others.

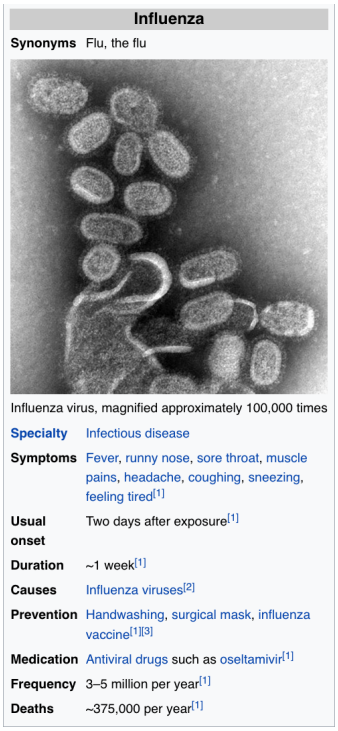


Figure 4: Infobox present in the English wikipedia page for influenza.

[2]https://en.wikipedia.org/wiki/Wikipedia:List_of_administrators
[3]https://pt.wikipedia.org/wiki/WikipÃĺdia:Administradores/Lista

Table 1: Median values for each metric by language, along with the p-value resulting from the Mann-Whitney test. Bold values represent the best values in each metric.

	EN	PT	*p*-value
Authority	**951.65**	708.5	<2.2e-16
Completeness	**2423.1**	576.8	<2.2e-16
Complexity	40	**72**	<2.2e-16
Informativeness	**0.53**	0.11	<2.2e-16
Consistency	**2582.04**	2073.63	<2.2e-16
Currency	**21**	245	<2.2e-16
Volatility	0.5	**0**	0.13

The goal of this extra set of measurements is, therefore, to compare the articles through the scope of some structural and content-related meta-data, in combination with the metrics which are primarily reliant on edit history rather than on content. These new measurements are: number of medicine templates, number of medicine infoboxes and number of citations.

4 RESULTS AND DISCUSSION

A statistical analysis was performed on each of the IQ metrics. As the samples of the several metrics in both languages do not follow a normal distribution, we have used the Mann-Whitney test to compare the means between the English and Portuguese articles.

The results of the test for each metric are shown in Table 1. The best results for each metric are displayed in bold. *Currency* and *Volatility* are the exceptions where a lower value is better, as lower *Currency* means more up-to-date articles and lower *Volatility* means faster recovery from vandalism.

According to the results presented in Table 1, we conclude that there are significant differences across most of the metrics between English and Portuguese medical articles in Wikipedia. The exception to this being *Volatility*, which means both Portuguese and English articles recover from vandalism in a similar time frame, and *Complexity*, which means Portuguese articles are easier to read than English articles.

One could argue that the *Complexity* metric is not adequate for this comparison. Martins et al. [10] states that the Portuguese Flesch adaptation was not researched extensively enough to guarantee complete accuracy, which can lead us to believe that the comparison between Portuguese and English article *Complexity* may not be reliable.

In comparison to the results obtained by Stvilia et al. [12], our results follow the same order of magnitude, with some inflation in metrics which involve a number of events in edit history, such as *Authority* and *Consistency*, which is to be expected given the date difference between the two analysis.

In Figure 5 we can see the number of medicine templates, medicine infoboxes and total citations in the English and Portuguese articles in our samples. Each data point in these graphics represents an article of a given language. These measurements were considered as particularly relevant in the context of medicine related articles. As one can observe from the figure, the English articles make more use of medicine related templates and infoboxes, as well as including a greater number of citations, when compared to Portuguese articles.

It should be noted that, from the English set, 129 articles were classified as *good* articles - articles which meet a set of editorial standards, but are not *featured* article quality - and 47 as featured, whereas in the Portuguese set only 15 articles were *good* articles and 27 were *featured*. English articles appear to use more templates and infoboxes, which may indicate a better effort in terms of content organization and reutilization of information, as well as more citations.

The benefits of English content found in this work are corroborated by others studies. Previously, Lopes et al. [9] have concluded that, when compared to Portuguese, English web content is more adequate for the distribution of health information.

5 CONCLUSION

We performed a quality comparison between English and Portuguese Wikipedia medicine articles, across a set of descriptive generic metrics, as well as some structural metadata and medicine-specific article elements.

We concluded that the medicine articles vary significantly in terms of quality, with the English Wikipedia outperforming the Portuguese Wikipedia across the majority the metrics used to describe the articles. Our results corroborate the importance of the investment in cross-language information retrieval strategies in the health domain.

As future work, we intend to investigate and develop a set of quality metrics specific to the medicine domain on Wikipedia, most likely using some of the extra measurements we used in this analysis.

ACKNOWLEDGMENTS

Project “NORTE-01-0145-FEDER-000016" (NanoSTIMA), financed by the North Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

REFERENCES

[1] 2018. WikiProject Medicine. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine
[2] 2019. The top 500 sites on the web.
[3] Joshua E. Blumenstock. 2008. Size matters: Word Count as a Measure of Quality on Wikipedia. *Proceeding of the 17th international conference on World Wide Web - WWW 08* (2008). https://doi.org/10.1145/1367497.1367673
[4] Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature* 438, 7070 (2005), 900–901. https://doi.org/10.1038/438900a
[5] James M Heilman and Andrew G West. 2015. Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language. *Journal of Medical Internet Research* 17, 3 (Apr 2015). https://doi.org/10.2196/jmir.4069
[6] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 453–462. https://doi.org/10.1145/1240624.1240698
[7] Jona KrÃĈÂđenbring, Tika Monzon Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. 2014. Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology. *PloS one* 9 (09 2014), e106930. https://doi.org/10.1371/journal.pone.0106930
[8] M. R. Laurent and T. J. Vickers. 2009. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association* 16, 4 (Jan 2009), 471–479. https://doi.org/10.1197/jamia.m3059

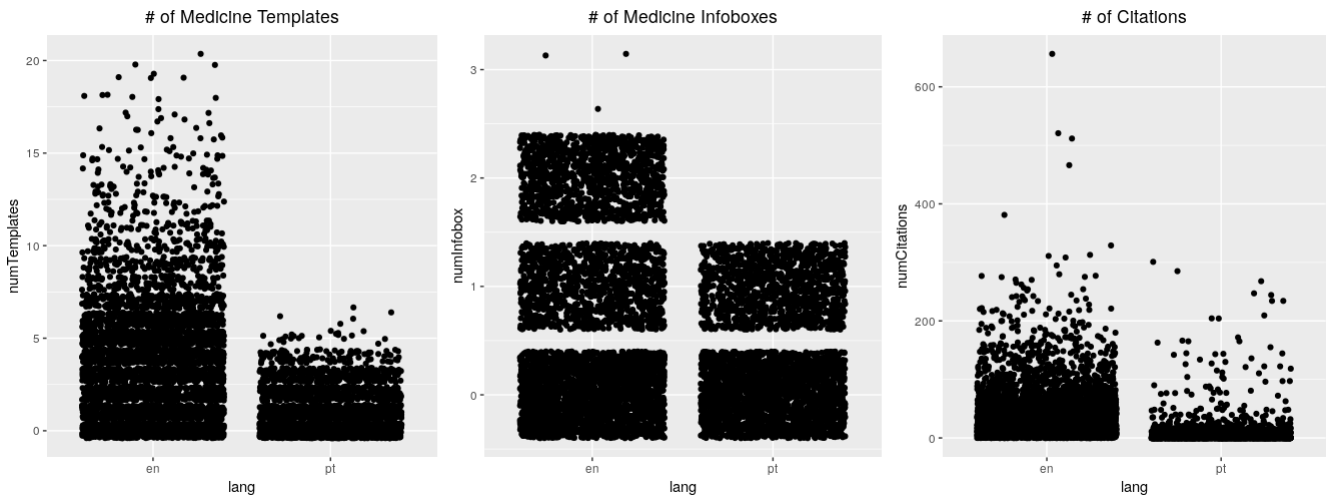


Figure 5: Number of medicine templates (left), medicine infoboxes (center) and citations (right) in the English and Portuguese articles.

[9] Carla Teixeira Lopes and Cristina Ribeiro. 2013. Measuring the value of health query translation: An analysis by user language proficiency. *Journal of the American Society for Information Science and Technology* 64, 5 (2013).

[10] Teresa B.F. Martins, Claudete M. Ghiraldelo, M. Graĝas V. Nunes, and O.N. Oliveira Jr. 1996. Readability Formulas Applied to Textboooks in Brazilian Portuguese. (Jun 1996).

[11] Garry R Thomas, Lawson Eng, Jacob De Wolff, and Samir Grover. 2013. An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology. *Seminars in dialysis* 26 (02 2013). https://doi.org/10.1111/sdi.12059

[12] Besiki Stvilia, Michael B. Twidale, Les Gasser, and Linda C. Smith. 2005. Information Quality In A Community-Based Encyclopedia. *Knowledge Management* (2005). https://doi.org/10.1142/9789812701527_0009

[13] Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. Mcguinness. 2006. Computing Trust from Revision History. (Jan 2006). https://doi.org/10.21236/ada454704

Thanks for Stopping By: A Study of "Thanks" Usage on Wikimedia

Swati Goel
ms.swati.goel@gmail.com
Henry M. Gunn High School

Ashton Anderson
ashton@cs.toronto.edu
University of Toronto

Leila Zia
leila@wikimedia.org
Wikimedia Foundation

ABSTRACT

The Thanks feature on Wikipedia, also known as "Thanks", is a tool with which editors can quickly and easily send one other positive feedback [1]. The aim of this project is to better understand this feature: its scope, the characteristics of a typical "Thanks" interaction, and the effects of receiving a thank on individual editors. We study the motivational impacts of "Thanks" because maintaining editor engagement is a central problem for crowdsourced repositories of knowledge such as Wikimedia. Our main findings are that most editors have not been exposed to the Thanks feature (meaning they have never given nor received a thank), thanks are typically sent upwards (from less experienced to more experienced editors), and receiving a thank is correlated with having high levels of editor engagement. Though the prevalence of "Thanks" usage varies by editor experience, the impact of receiving a thank seems mostly consistent for all users. We empirically demonstrate that receiving a thank has a strong positive effect on short-term editor activity across the board and provide preliminary evidence that thanks could compound to have long-term effects as well.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing*.

KEYWORDS

Wikipedia, Motivation factors, Thanks feature

ACM Reference Format:
Swati Goel, Ashton Anderson, and Leila Zia. 2019. Thanks for Stopping By: A Study of "Thanks" Usage on Wikimedia. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.3316756

1 INTRODUCTION

Wikipedia is an online encyclopedia that functions largely because of volunteer editors. Editor engagement and motivation are therefore central to Wikipedia's existence. Research has long shown that positive external motivation (rewards, recognition) can lead to an increase in contribution to a community, and it is well-established that people will contribute more to a group the more enjoyable they find it [4]. A positive community environment, which provides editors a better experience, can therefore increase editor activity. A positive environment may actually be one of the most crucial elements for increasing engagement, as social factors tend to outweigh even those surrounding usability with regards to positively affecting contribution [3]. The impact of these social factors could be quite significant, as a community member's internal value systems can be influenced by external rewards, thus making positive feedback an extremely useful tool in building online communities [6]. The Thanks feature could therefore represent an important resource for building a positive Wiki community.

"Thanks" is no longer a new Wiki feature, having been implemented on English Wikipedia on May 30th, 2013 and introduced to all projects soon thereafter. In contrast to previously existing features such as WikiLove, it allows editors to thank one another for *specific* revisions in a *semi-private* way. There is a public record of who thanked whom and when, but the specific revision and the article for which a user was thanked is private information. This setup is quite different from that of previous community building features such as WikiLove, yet the literature on "Thanks" is limited. Previous research on "Thanks" includes a study by Harburg and Matias [2], which draws an interesting contrast between the Thanks feature and WikiLove. We do not interact directly with their work as their main focus is analyzing usage differences between the two features whereas we study the impact of "Thanks" as well as its usage more deeply. Part of our work builds upon the thoughtful cross-cultural analysis of Nemoto and Okada [5], which examines differences in "Thanks" usage across languages. In addition, we conduct more in-depth research on how "Thanks" is used over different levels of editor experience and how thanks are distributed over time. We also examine the Thanks feature's impact on editor engagement, controlling for editor experience between comparisons of thanked and unthanked editors, and we find that thanks are linked to increases in short-term activity, in contrast to the results of the Nemoto-Okada study. This is significant because, in conjunction with our findings of the Thanks feature's limited scope, our results suggest that "Thanks" has additional potential to increase editor motivation.

2 METHODOLOGY AND RESULTS

Our work is separated into two parts: analyzing how people interact with "Thanks" and assessing the impact the feature has on editor engagement and motivation. In part one, we present data on general use, such as the characteristics of the feature's users, how likely different types of editors are to receive thanks, etc. We then establish a correlation between receiving a thank and having a higher edit count. This naturally leads to part two, in which we attempt to determine whether this link is causal. We discuss our results at a high-level below.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3316756

Language	Thanks Givers	Thanks Receivers	Editors	% Thanks Givers	% Thanks Receivers
German	23433	31567	390603	6.0	8.08
Español	14079	13924	526009	2.68	2.65
Italian	8742	9412	186733	4.68	5.04
Portuguese	8093	8593	194509	4.16	4.42
Polish	5880	6506	83949	7.0	7.75
Farsi	4611	4829	108114	4.26	4.47
Dutch	4704	4830	89006	5.29	5.43
Arabic	6873	6662	148112	4.64	4.5
Korean	1855	1874	68285	2.72	2.74
Thai	1045	610	29780	3.51	2.05
Norwegian	1171	2500	41501	2.82	6.02

Table 1: Scope of Thanks feature in select languages. Each column represents the absolute count of editors in a category unless explicitly specified as a percentage. English is excluded due to data restrictions.

2.1 How "Thanks" is Used

2.1.1 Scope of "Thanks". The scope of the Thanks feature, or the number of editors who have either given or received a thank since the feature was first introduced, is generally between 4-6% (in a subset of larger languages), as shown in Table 1. In the set of editors with 5+ edits, the scope of the feature is 15-17% (again in a subset of larger languages), indicating the existence of a small group of active editors who are responsible for a vast majority of thanks.

The Thanks feature is not as widely used as it could be, but it has become more prevalent in recent years, a trend shown in Table 2. Even in languages where the absolute editor count has dropped, "Thanks" usage rates have increased. There is a clear upward trend in the number of thanks givers (relative to the number of total editors). Somewhat surprisingly, this same trend does not hold for thanks receivers. The generally increasing usage rates for givers suggest that some editors are only now beginning to use the Thanks feature, which could indicate that a large portion of the editor population remains unaware of it. If this is the case, efforts to increase exposure could have significant benefits, especially because novice editors—those for whom the feature could potentially be the most valuable—are currently the ones who interact with it the least.

2.1.2 Usage by Editor Experience. The distribution of thanks shows that novice editors interact with the Thanks feature less frequently than their more experienced counterparts. Presented in Table 3 is the average number of thanks received by a set of editors (grouped by editor experience) per month or day counting only months or days in which they received at least one thank. We do not include editors who have never received a thank, and we partition the remaining data into novices (bottom 20% of editors by edit count) and experienced (top 20% of editors by edit count). The editors we studied received multiple thanks on the same day more often than they would have if thanks were given at random times, implying a non-uniform distribution, and there was a strong correlation between those who were given more thanks and those who had higher edit counts. These two findings taken together indicate that thanks are often received in 'clumps' and awarded to those who are editing more frequently. This more concentrated "Thanks" usage may be why the feature has the outsized impact shown in part two, but it's also possible that a more widely used feature would retain the same benefits while reaching more editors.

The existence of a positive thanks to edit count correlation is further corroborated by our analysis of average thanks given over all editor percentiles: the top 5% of editors give by far the most thanks in absolute terms. However, as Figure 1 shows, these editors give the least thanks relative to their edit counts.

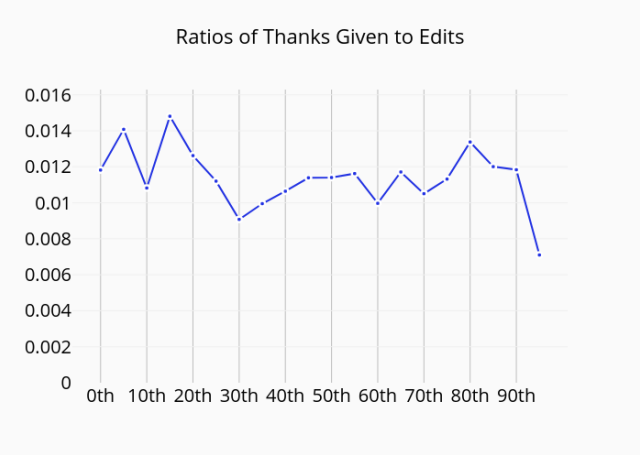


Figure 1: "Thanks" by edit count ratios. Percentile are based on edit count. Data from June '17-'18

2.1.3 Additional Work. The research page for this project [7] contains links to a number of sub-projects left out of this report as well as code pipelines for anyone interested in replicating the project. A brief summary of some of these projects follows:

- In order to better understand the scope of "Thanks", we define different levels of editor engagement with the feature. We calculate the percentage of total editors who contributed to random samplings of thanks and find that just under 5% of editors in our dataset (a group of larger Wikipedia languages) were responsible for giving 80% of thanks.

Language	Givers '18	Givers '16	% Givers '18	% Givers '16	Receivers '18	Receivers '16	% Receivers '18	% Receivers '16
Italian	1910	1511	6.03	4.98	2195	1995	6.93	6.57
Portuguese	1273	1314	5.16	4.75	1835	1601	7.44	5.79
Polish	1297	940	8.67	6.66	1349	1419	9.02	10.06
Farsi	1103	575	5.72	4.47	927	789	4.81	6.13
Netherlandic	935	915	6.82	6.28	1046	1096	7.63	7.52

Table 2: "Thanks" usage rates. Each column represents the absolute count of editors in a category unless explicitly specified as a percentage. Data is from the first 6 months of each year.

Language	Sample	Year	Month	Day
Italian	Bottom 20%	2.69	1.68	1.18
Italian	Top 20%	119.62	13.07	1.59
Portuguese	Bottom 20%	2.95	1.98	1.34
Portuguese	Top 20%	206.24	22.22	2.33
Polish	Bottom 20%	2.34	1.63	1.19
Polish	Top 20%	48.63	6.3	1.42
Farsi	Bottom 20%	2.73	1.91	1.28
Farsi	Top 20%	123.0	13.74	1.74
Dutch	Bottom 20%	2.37	1.48	1.11
Dutch	Top 20%	81.0	9.61	1.5

Table 3: Distribution of thanks for individual editors. Each number represents the average number of thanks received by an editor in a timeframe considering only those in which at least one thank was received.

- To characterize thanks senders vs thanks receivers, we compare the average tenure and edit count of the former group to the average tenure and edit count of the latter group and find that in most languages studied—Norwegian being a notable exception—thanks are generally sent from newer, less experienced editors to more experienced, veteran editors. This trend is further corroborated by Nemoto and Okada.
- To get a sense of the variance in "Thanks" usage across projects, we rank all Wikimedia projects at the time of study (July 2018) by the ratio of editors who had sent a thank to the total number of editors since "Thanks" was introduced.

2.2 Editor Engagement

In part two, we establish a link between receiving a thank and having a higher future edit count, at least in the short-term. To do this, we match editors who received a thank on some day with editors who did not receive a thank on some (potentially different) day and compare their subsequent edit activity. Because we only match between editors with similar characteristics, we can be reasonably confident that the thank, and not some other factor, causes the future edit count differences we see. Table 4 presents a result of the study in Polish Wikipedia (trials in Portuguese Wikipedia and MetaWiki yielded similar results). In the table, the feature data for each cohort (ex: tenure for thanked editors) is an average of that feature's value over all members of that group. The 'editors with higher counts' field represents the number of editors of the group who had a higher subsequent edit count than their match.

We find a positive correlation between the five features we use (tenure, edits, thanks, short-term edits, short-term thanks) and the dependent variable, future edit count. In Table 4, unthanked editors have a higher average value for every feature. We would therefore expect them to have a higher average future edit count as well. This is not reflected in the data, suggesting that the difference we see in future edit count is caused by the one field along which the groups are not balanced: whether or not an editor has just been thanked. It is possible that the edit count discrepancy is actually caused by some unaccounted for confounding variable, but our test results suggest this is not the case. A random forest classifier we trained, for example, demonstrated that the features we chose have good predictive accuracy (a little over 90%). Using both the features and a field for whether the treatment of receiving a thank was applied, the classifier accurately determined the range in which future edit count would fall. This suggests that our features were comprehensive. Additionally, another test revealed that a variety of random features either had less weight on the overall prediction than our selected five or led to similar matchings.

3 CONCLUSIONS

"Thanks" usage rates have been increasing over time, even in projects where the absolute editor count has dropped. This could mean that current low "Thanks" usage rates among novice editors are due to editors being unaware of the feature as opposed to them being opposed to it, and it indicates that increasing the Thanks feature's exposure could increase usage. Currently, thanks seem to be centered around a few editors and are not sent as a matter of course, which could change with increased attention to the feature. More attention might also alter the "Thanks" social structure in which most thanks are sent from less experienced to more experienced editors. Having more positive feedback go from experienced editors to novices seems intuitively likely to encourage newer editors to stay involved, and in fact, we show that "Thanks" is linked to higher short-term editor activity. Because the feature is largely used by a group of editors who are already active and committed, each individual thank is unlikely to have more than a short-term impact. However, given our results, it's conceivable that the effects of thanks may compound over time and that receiving a thank as a novice editor could change a Wikimedian's career. While the long-term effects of thanks were not determined in this study, our findings suggest that increasing "Thanks" usage would positively affect editor retention and activity. We would not want thanks to become so common as to be meaningless, but the feature is far from that point, if it exists. Thus, it is our belief that connecting

Group	Tenure	Edits	Thanks	Short-term Edits	Short-term Thanks	Next Day's Edits	Editors with Higher Counts
Thanked	2848.3	555.3	5.8	56.8	0.3	9.6	46
Unthanked	3021.7	573.1	5.9	60.6	0.3	5.7	15

Table 4: Motivation study results sample

more editors with the Thanks feature would be beneficial to editor motivation and possibly editor retention.

REFERENCES

[1] Fabrice Florin. 2018. Information for Thanks. (2018). https://en.wikipedia.org/wiki/Wikipedia:Notifications/Thanks

[2] Emily Harburg and J Nathan Matias. 2014. The Effect of Appreciation on Wikipedians. (2014). https://civic.mit.edu/2014/11/16/researching-love-and-thanks-on-wikipedia-crowdcamp-hackathon-report/

[3] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to participate in online communities. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden (Eds.). ACM, 1927–1936. https://doi.org/10.1145/1753326.1753616

[4] Kimberly S. Ling, Gerard Beenen, Pamela J. Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley, Dan Frankowski, Loren G. Terveen, Al Mamunur Rashid, Paul Resnick, and Robert E. Kraut. 2005. Using Social Psychology to Motivate Contributions to Online Communities. *J. Computer-Mediated Communication* 10, 4 (2005). https://doi.org/10.1111/j.1083-6101.2005.tb00273.x

[5] Keiichi Nemoto and Ken-ichi Okada. 2015. WIKI THANKS: Cultural Differences in Thanks Network of Different-Language Wikipedias. *CoRR* abs/1502.04312 (2015). arXiv:1502.04312 http://arxiv.org/abs/1502.04312

[6] Steven J. J. Tedjamulia, Douglas L. Dean, David R. Olsen, and Conan C. Albrecht. 2005. Motivating Content Contributions to Online Communities: Toward a More Comprehensive Theory. In *38th Hawaii International Conference on System Sciences (HICSS-38 2005), CD-ROM / Abstracts Proceedings, 3-6 January 2005, Big Island, HI, USA*. IEEE Computer Society. https://doi.org/10.1109/HICSS.2005.444

[7] Leila Zia, Swati Goel, and Ashton Anderson. 2018. Understanding Thanks. (2018). https://meta.wikimedia.org/wiki/Research:Understanding_thanks

Understanding Travel from Web Queries using Domain Knowledge from Wikipedia

Chander J Iyer
chander.iyer@verizonmedia.com
Yahoo! Research
Sunnyvale, California

Srinath Ravindran
rsrinath@verizonmedia.com
Yahoo! Research
Sunnyvale, California

ABSTRACT

Developing a deeper understanding of the travel domain is helpful for presenting users with consistent and reliable information, and few sources of data are able to achieve that. Further, such information can serve as background knowledge for evaluating machine learning algorithms. In this paper, we present part of our work towards developing such an understanding. We demonstrate a simple extraction technique and how the extracted data can be used to evaluate an unsupervised embedding model built on search queries with travel intent.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; • **Information systems** → *Language models*; Wikis.

KEYWORDS

Wikipedia; label generation; model evaluation; query embedding

ACM Reference Format:
Chander J Iyer and Srinath Ravindran. 2019. Understanding Travel from Web Queries using Domain Knowledge from Wikipedia. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3316759

1 INTRODUCTION

Planning a trip is one of the many stressful things we do everyday, and lack of a good source of information is one of the major reasons. Travel typically constitutes a trip to a destination including the mode of transport to the destination, the means of stay and commute within the destination. Various websites such as Tripadvisor, Expedia, and other travel blogs provide the reviews and details about hotels, flights, and car reservations. However, they either (a) lack consistency in tiers of travel, (b) promote sponsored content, or (c) contain outdated information.

In this paper, we present a part of our work which focuses on developing a deeper understanding of travel, and present how such a knowledge can be used to (a) present users with consistent and reliable information, and (b) evaluate machine learning algorithms.

Knowledge extraction tasks based on Wikipedia come in a few flavors: training embedding models [1], building knowledge graphs and ontologies [6], or natural language processing and extraction [3]. Existing knowledge bases like YAGO or DBPedia that build information extraction pipelines using Wikipedia [7] require significant effort to editorially curate the extracted data and label data for training models. Unlike other work in the past, we do not intend to build a knowledge graph or an ontology. Instead, we utilize data from Wikipedia directly. Lots of data are required to build a large ontology or to develop sophisticated machine learned techniques to parse and extract the information from web data. As we shall demonstrate, most of the information can be obtained from Wikipedia using a simple extraction process with almost no natural language parsing. Further, one of the arguments against an unsupervised model is the lack of data, and our approach demonstrates how the representation of data of Wikipedia enabled us to evaluate a machine learning model.

2 KNOWLEDGE IN WIKIPEDIA

In our work, we are interested in extracting specific Wikipedia entities associated with hospitality and travel along with relevant metadata. In particular, we extracted the brand name or the company name, the name of the establishments, their tier of service, and the location. For a hotel, that would mean, we extract the parent company, say 'Wyndham', and then the list of establishments owned by the parent such as 'Days Inn', 'La Quinta', 'Ramada', 'Super 8' and 'Wyndham Grand'. For each of these, we extract their tier of service such as 'upscale', 'mid scale', 'boutique', and 'economy'.

2.1 Information Extraction

The information we seek is predominantly distributed across various organized parts of Wikipedia such as lists, sections, categories, info boxes and templates.

We start with the Wikipedia page: 'List_of_lists_of_lists', which points to several other lists. For the purposes of the work described in this paper, we focused on travel within United States to include list of hotels, airports and car rental companies. In addition to the lists, we also used the Category pages such as 'Category:Vehicle_rental_companies' and 'Category:Travel_and_tourism_templates' as secondary starting points, for topics not listed in the lists. These pages yield the company names and the brand names.

Next, we derive info such as the tier of service and the locations, where appropriate, using the sections, info boxes and subcategories within the pages for each of the brands we extracted.

Finally, we sought human editorial help[1] for verifying various labels such as tiers and categories for the extracted data. Remarkably, the travel activities annotated using Wikipedia extractions

[1] an in-house editorial team at Verizon Media

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316759

agreed with editorial review over 65% of the time. The accuracy was lower than expected since the brand differences are subtle and can fall into multiple tiers of service e.g. upper midscale hotels can be classified as upscale or midscale. This reinforced our belief that data from Wikipedia is sufficiently accurate, and some of gaps in agreement were due to standardization of labels.

2.2 Challenges

While we highlight the ease of information extraction from Wikipedia, it is important to point out the challenges and limitations we encountered.

- **Incomplete Data** While Wikipedia continues to grow daily, the data is not complete. For instance, not all hotels in a city are listed, and lesser known establishments are often missing.
- **Inferring Complex Relationships** Let us consider the example of airports. In the current state, it is not possible to derive information such as nearby airports without advanced processing or potentially building an ontology. While sites like Wikivoyage support complex queries to extract nearby airports, they are limited by the Wikipedia ontology of neighboring attractions to extract nearby airports.
- **Lack of Consistency within Categories and Templates** Across several pages within Wikipedia, the categorization and tiers of service labels are inconsistent and there is some disagreement about the labels, and we have to normalize these labels.
- **Choice of labels for entity categorization** Vehicle rentals could be categorized as car rentals, truck rentals, rideshares etc. or they could be categorized by car design size such as Compact, Mid-size, Family etc. Choosing the appropriate set of Wikipedia labels for categorization is a well-known problem [4].
- **Lack of Consistent formatting within Wikipedia** Not all lists are formatted the same. While some are organized as tables, others are enumerated or are bulleted lists. Parsing such pages takes some effort.

In spite of these shortcomings, we find that the data in Wikipedia is a good starting point for large scale information extraction. Since they cover the most frequent or most popular destinations, commonly referred to as the head, the data is good for initial prototypes.

3 CASE STUDY: UNDERSTANDING TRAVEL SEARCH QUERIES

We studied a time ordered trail of search queries of users from Yahoo! Search logs, with the intent of understanding typical travel related search patterns within user search sessions. The data used was anonymized. Queries such as "cheap flight tickets to Orlando" or "Sheraton hotel address in downtown Atlanta" and "Enterprise rental Dallas airport" are examples of typical queries with travel intent. We built an embedding model to derive the relationships between search queries, and we used data extracted from Wikipedia to evaluate our model. We describe each of the process below.

Table 1: Geographical and Brand-Tier similarity scores for all travel activities

	NN=1	NN=2	NN=3	NN=5	NN=10
Geographical	0.536	0.502	0.479	0.446	0.403
Brand − Tier	0.552	0.528	0.514	0.498	0.476

Table 2: Geographical and Brand-Tier similarity scores for vehicle rental activities

	NN=1	NN=2	NN=3	NN=5	NN=10
Geographical	0.583	0.534	0.506	0.465	0.415
Brand − Tier	0.633	0.600	0.569	0.523	0.462

3.1 Query Embedding

We train an embedding model similar to Word2Vec[5], where users are equivalent to a document, search queries are the 'words' and a search session is a 'sentence'. Similar models have been proposed in the past for query understanding [2] to generate a 300-dimensional embedding for each search query. In the resulting embedding space, related travel activities appear closer to each other. We extract the top 100 neighbors of each travel activity using cosine similarity for each travel activity pair. We set the threshold frequency for travel queries to have at least $f = 500$ occurrences for generating the activity embeddings.

3.2 Query Evaluation

For each travel activity query, we compute the similarity scores for the top k nearest neighbors where $k = \{1, 2, 3, 5, 10\}$. For our case study, we computed similarity scores across 2 dimensions:

(1) Geographical similarity
(2) Brand-Tier similarity.

We augment a given travel query using the domain knowledge from Wikipedia sections that were extracted using the process described earlier in Section 2.1. The augmented data is then used to compute the similarity scores. For example, a traveler searching for "cheap flight tickets to Orlando" could also be potentially interested in "budget hotels near Disney Orlando" or "economy car rentals in Orlando airport". We expect such queries to be more related to each other and other queries related to Disney than a query like "hotels near Aspen ski resort".

3.3 Results

Table 1 shows both geographical (city or state) similarity and the brand level or category level similarity scores across measured on a set of 20k activities for the top nearest neighbors. An $NN = 1$ score of 0.536 for the *Geographical* score indicates that 53.6% of the first-nearest neighbors of queries had the same geographical intent (same city). The similarity scores decrease with the increase in the number of neighbors. This is consistent with the fact that the farther neighbors are less relevant than the nearer ones.

As observed, the *Brand–Tier* scores are better than *Geographical* similarity scores across all travel activities. One reason for poor *Geographical* scores is while flights and cars are reserved at specific city airports, the corresponding hotel reservations could be at a nearby city suburb where the intended travel activity is performed e.g. while a flight reservation is made for Dulles airport in Washington DC, the hotel reservation could be made in Alexandria, Virginia. This effect is more pronounced for specific activities like vehicle rentals as shown in Table 2. On the other hand flight reservations exhibit better geographical similarities than other activities.

4 CONCLUSION

In this paper, we presented a part of our work in progress to that demonstrates the use of Wikipedia to develop a deeper understanding of the travel domain. In particular, we described an approach to extract entities and their labels, and showed how to use the extracted information to evaluate an embedding model. Subsequent work is aimed at improving the extractions to increase of the depth of knowledge. We are also exploring ways to solve some of the challenges described in Section 2.2.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* abs/1607.04606 (2016). arXiv:1607.04606 http://arxiv.org/abs/1607.04606

[2] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. 2015. Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 383–392. https://doi.org/10.1145/2766462.2767709

[3] Lili Kotlerman, Zemer Avital, Ido Dagan, Amnon Lotan, and Ofer Weintraub. 2011. A Support Tool for Deriving Domain Taxonomies from Wikipedia. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Association for Computational Linguistics, 503–508. http://aclweb.org/anthology/R11-1069

[4] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 115–124. https://doi.org/10.1145/3077136.3080834

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

[6] Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a Large Scale Taxonomy from Wikipedia. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2 (AAAI'07)*. AAAI Press, 1440–1445. http://dl.acm.org/citation.cfm?id=1619797.1619876

[7] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. https://doi.org/10.1145/2629489

What's in the Content of Wikipedia's Article for Deletion Discussions?

Towards a Visual Analytic Approach

Ali Javanmardi
Department of Computer Science
University of Western Ontario,
London, ON, Canada,
sunny_javanmardi@yahoo.com

Lu Xiao†
School of Information Studies
Syracuse University, Syracuse, NY,
U.S.A., lxiao04@syr.edu

Keywords

Wikipedia discussions; Wikipedia Article for Deletion dataset; Visual analytics; Natural language processing

ACM Reference format:

Ali Javanmardi and Lu Xiao. 2019. What's in the Content of Wikipedia's Article for Deletion Discussions?: Towards a Visual Analytic Approach. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316750

1 Introduction

As a successful decentralized online peer production system, Wikipedia has a large number of users participating in self-assigned tasks based on their interest. Participants often coordinate on the tasks about editing an article through the article's talk page discussions. Researchers analyze Wikipedia talk page discussions for various aspects such as coordination [13], conflict management [1], emotion [8], leadership [12], the relationship between Wikipedia article's talk page discussion and the article editing activity [5], etc.

Besides talk page discussions, Wikipedia's deletion discussions have also caught researchers' attentions. In Wikipedia, there are four mechanisms with respect to deleting an article from the web site: speedy deletion, proposed deletion, BLP Deletion, and deletion discussion. Deletion discussion mechanism is designed for controversial articles that the deletion needs to be discussed to reach a conclusion. Also, if a proposed article cannot be tagged with one of the first three mechanisms, it will be tagged for deletion discussion.

Researchers have explored various factors of the deletion discussion process and outcome. For example, [19] examined voting behavior in deletion discussions, with a specific emphasis on how early opinions in the debate affect later voices. [6] discussed the importance of being familiar with the Wikipedia community's values, norms, and practices in deciding whether the discussed article should be deleted. [9] considered the impact of Wikipedia's "Ignore All Rules" policy on decision-making in deletion debates. [21] found that the outcome of an article's deletion discussion correlates with the type of votes in the discussion and the category of the article's topic (e.g., art, location, sport, etc.).

Through open and iterative coding processes, researchers analyzed the deletion discussion content. They found that these discussions are logical and follow Wikipedia's policies in general, and notability is the most commonly used rationale for keeping or deleting an article [14, 21]. The discussion participants also referred to various Wikipedia policies and guidelines in justifying their positions regarding the articles [21, 22]. Nonetheless, these studies analyzed small samples, which made it challenging to generalize their findings. Analyzing a large amount of the discussion content is desired to help better understand and leverage this discussion space, but requires computational content analysis. On the other hand, the deletion discussions are unstructured text data. Significant efforts are needed to clean raw discussion content for computational analysis. For example, a user may forget to add a closing tag or his/her signature at the end of the comment, which makes it difficult to detect the end of a user's comment.

To address this challenge, we retrieved about 40,000 deletion discussion content from Wikipedia web pages, cleaned and stored the content of 39, 177 discussions into a structured discussion database. With this cleaned and structured dataset, the automatic processing and analysis of the discussion content becomes more manageable to the researchers. With this database, we developed interactive visualizations that offer insights on how the outcomes of the articles are related to different aspects of the discussions, including the types of votes, the mentioned policies, and the categories of the articles (e.g., art, people, sport, etc.). This database is available at https://drive.google.com/open?id=1n3GJ6h5bSykrC7cgE2_oz4N6LSM3UnaV.

The rest of the paper is organized as follows. We first review studies related to the factors of the deletion discussions and visualizations. We then describe in details the open access structured database including how we cleaned and filtered the noise, and annotated and categorized the data. We then present our interactive visualizations and discuss their implications. We

† Corresponding author.
This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316750

conclude with new research ideas that leverage the open access database, shedding light on the future research direction.

2 Related Work

2.1 The Decision Factors in Wikipedia Deletion Discussions

The deletion discussions in Wikipedia are often referred to as Article for Deletion (AfD) discussions. An AfD discussion starts when a user nominates the article to be discussed for deletion and offers his/her justification for the nomination. During the discussion time, any Internet user, often a Wikipedian, can participate in the discussion and deliberate on what to do with this Wikipedia article - to delete it from the Wikipedia, to keep it, or to make another strategic move such as merging it to another article. The participants are required to provide their opinions as well as the rationales that justify their opinions. As a main quality control mechanism, the users' opinions and rationales are expected to follow Wikipedia's policies and requirements about the articles. When a Wikipedia administrator or a general user makes the decision about the article, the AfD decision should be based on the rationales provided by the participants, instead of the majority vote rule of the participants' opinions. Also, often the user who makes the decision and closes the discussion is someone who did not participate in the discussion process.

AfD discussions are mainly con-ducted by a small number of long-established users [6, 18]. The early opinions in a discussion are found to affect the following opinions, and the users exhibit different voting patterns in the process at subgroup level [19]. The naturally formed groups make better decisions than those that recruit participants, and a group with a moderate diversity of newcomer and expert participants concludes better decision [11]. Additionally, non-unanimous situations tend to have more participants, which implies that participants are more inclined to voice their different opinions than agreeing to others 14, 21].

[21] examined whether and how different types of votes affects the outcome of an AfD discussion. They found that the existence of other vote (i.e., a vote other than keep or delete) has a statistically significant impact on the outcome of the article – an article that has the other votes in its deletion discussions is more likely to be suggested for actions other than delete. The authors also found that certain categories of an article correlates with the likelihood of the article to be deleted. For example, articles about people, for-profit organizations, or definitions are slightly more likely to be deleted than expected, while articles about locations or events are more likely to be kept than expected, and articles about nonprofit organizations and media are more likely to be suggested for other options (e.g., merge, redirect, etc.) than expected.

"On Wikipedia, notability is a test used by editors to decide whether a given topic warrants its own article" (https://en.wikipedia.org/wiki/Wikipedia:Notability). Prior studies found that notability policy indeed plays a critical role in the AfD discussions. Taking the assumption that an article's level of popularity correlates well with its level of notability, [10] measured an article's notability by measuring its readership (i.e., the number of visits to the article) and its rank in the Google search. The authors found that the most common reason for deleting an article is its lack of notability. A later content analysis study (N = 229 AfD discussions) also found that notability is the most commonly used rationale for keeping or deleting an article [21].

While these studies have shown interesting results, their data samples are limited, especially in those that focused on the content analysis, e.g., [14, 21].

2.2 Visualizations with Wikipedia Data

Wikipedia data have been used in various visualization projects. For example, 509 administrators' activities in Wikipedia's article editing process were visualized into a colorful tabular format (Chromogram) which showed that the administrators switch between multiple tasks [20]. [2] implemented a radial layout visualization to illustrate the relationships among the categories of Wikipedia. They also visualized data streams of recent changes in Wikipedia by mapping time and activity levels into a two-dimensional display of x and y respectively, and encoding the different types of activities to different shapes [3]. For more examples of visualizations based on Wikipedia content, please refer to the web site: http://infodisiac.com/Wikimedia/Visualizations/.

[17] and [19] leveraged visualizations to investigate Wikipedia's AfD content. The authors chose 100 longest AfDs that had the deletion outcome and visualized them based on the type of the votes (keep or delete) and the sequence of the votes. The researchers also chose 100 longest AfDs that had the keep outcome and visualized them using the same strategy. In addition, they visualized the frequency distribution of the AfDs regarding the length of the discussions. They showed that a visualization can elaborate more levels of details of an online deliberation to the Wikipedia community.

3 An Open SQL Database for AFD Study

AfD discussions are informal and unstructured just as the typical online discussions. It is a technical challenge to apply computational techniques to analyze this type of texts – there are human errors (e.g., typos and poor grammar) and noises (e.g., mis-format, Internet slang). We detail how we obtained, parsed, and organized a subset of Wikipedia AfD data into a relational database that is ready for computational processing or analysis. It is available at https://drive.google.com/open?id=1n3GJ6h5bSykrC7cgE2_oz4N6LSM3UnaV.

Wikipedia manages the deletion discussions based on the date they are proposed for Deletion Discussion. The content of each proposed date is publicly accessible through URLs. An AfD discussion consists of four parts: article title, nomination reason – why the article is proposed for deletion, participants' comments – the participant's opinion and rationale to justify the opinion, and discussion outcome – the decision regarding the article and the rationale of this decision. With the retrieved data,

we observed over thirty types of opinions, e.g., weak delete, delete, withdraw, userfy, speedy decline, relisted, move, etc. And, there are five possible outcomes: Delete, Keep, Redirected, Merged, or None-consensus. All these parts except the first one end with user signatures. A user signature consists of the user's username and timestamp (date and time). Our script visited the web sites through the URLs of the proposed dates from May 15, 2013, to May 15, 2015 (i.e., 720 dates) and extracted these parts from the raw HTML content. To prepare the dataset for further computational analysis techniques, we removed 5% of the dataset that caused difficulties when parsing the HTML content to extract the parts. These difficulties come from the following noises:

Fatal Errors Fatal errors cause a complete shutdown in the parsing process. When such errors occur, we deleted the whole AfD discussion. On average, every 275 AfDs generates 3 to 4 such fatal errors, that is, 1.1% to 1.5% of the error rate.

The Error of Closing and Opening Tags Occasionally, AfD participants did not close HTML tags in their comments. Thus, there are opening tags without the closing tags. These HTML tags are often nested making it more difficult to detect the missing spot automatically. About 0.7% of the AfDs had this problem.

Noises over Very Long Discussions The longer an AfD discussion content is, the more complex it is to handle the noise. In some cases, the first author manually went through the AfDs to tackle the noise. The error rate for this kind is 1%.

Noises over User Signatures Whenever a user participates in an AfD discussion, his signature is automatically generated by the system. Each signature consists of the username and the timestamp. Technically speaking, a user's signature can used to separate subparts of an AfD from each other. However, the user signature has noises. For example, a timestamp might be completely missing from the end of a signature. The error rate is between 0.8% and 1.4% for the noise due to the user signature.

Through this cleaning process, we stored 39, 177 discussions in the SQL database from the raw about 40, 000 discussions. Approximately 80% AfD discussions had between 4 and 12 comments. Additionally, we stored different aspects of the discussion information into different tables and fields. Besides the aforementioned four parts, we also calculated and recorded other information, e.g., the total number of comments in an AfD discussion as well as the total number of keep/delete opinions it contains. Additionally, we identified and stored the category of a Wikipedia article and the Wikipedia policies mentioned in the discussion, as well as the hierarchical relationships among different categories and among different policies. We detail these processes below.

The Category of a Wikipedia Article Wikipedia offers a semi-hierarchical structure scheme for article categories such that one category may have multiple parents besides multiple children. In other words, Wikipedia has a category graph not a tree. Converting this graph into a tree is not a trivial task. To tackle this problem, [3] first constructed the graph, and then identified the parent category for each category by measuring the category similarity based on the co-occurrences of the articles in the two categories and the depth of the current category in the category graph. [4] developed a technique to address the noise in the hierarchy level of the Wikipedia article categories. In our SQL database building process, the first author compared all the parent categories of a category and chose one to be the parent category based on his judgment on the similarity between the parent category and the article being labelled.

The Policy Structure in the Database A Wikipedia policy can have a parent policy and one or more children policies. For example, the notability in a politician is the sub-set of the notability policy. The retrieved HTML data from AfD discussion pages do not show the parent-child relationships of the policies. Additionally, the participants often used shortcuts of a specific policy in the discussion. The mapping between a policy and its shortcuts is usually one-to-many and such knowledge is nested in Wikipedia pages, i.e., there isn't one place where such mapping is presented comprehensively for a policy. For example, Wikipedia:Deletion policy has three shortcuts WP:DP, WP:DEL, and WP:DELETE listed in the directory. But when one goes to the actual policy page, there are more than 20 additional shortcuts for different parts of the policy description. To tackle this challenge, the first author browsed all the Wikipedia policy pages to identify these inner-relationships and the use of multiple terms and shortcuts, and entered the data to the policy table manually. It is worth noting that while Wikipedia site distinguishes a policy from its guideline essays, many guideline essays are referred to and regarded as the policies in the community. Our database therefore did not distinguish the guideline essays from their policies.

Through this process of labelling categories and extracting policy information from participants' comments, the database has about 86% of the comments referred to Wikipedia policies. The total number of unique categories and policies are 321 and 3,583 respectively.

4 Factors of an AfD Discussion Outcome

Taking a visualization approach to explore the factors of an AfD discussion outcome, we developed interactive visualizations to obtain an overall understanding of how the outcomes of the articles correlate with different aspects of the discussions. In our visualizations, we coded a participant's opinion as follows. Code *delete* includes opinions like strongly delete, speedy delete, delete, and weak delete. Code *keep* includes the opposite opinions or final decisions (e.g., weak keep, strong keep, etc.). All the other types of opinions are coded as *other*. In addition, we only considered the discussions that had one of the two discussion outcomes: *keep* (the article) vs. *delete* (the article).

4.1 The Outcomes vs the Number of Keep/Delete Opinions

According to Wikipedia's AfD policy, the decision regarding a proposed article should be based on the participants' rationales in their votes, as opposed to simply following the majority vote based on the total number of different opinions in the discussion.

Curious to find out whether the actual decision-making of AfD discussions follows this policy, we visualized the AfD discussion outcomes and the number of keep/delete opinions in the discussions through an interactive heat map. In the remaining session, we call these opinions votes for the ease of comparison with the idea of majority vote. This interactive visualization is accessible at http://www.mandanemedia.com/afd/view/diagram1.php. Figure 1 shows a screenshot.

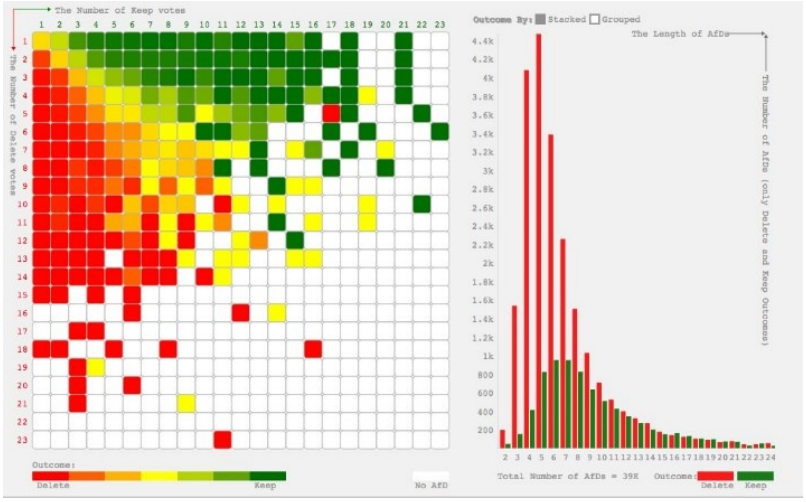


Figure 1: The number of keep and delete votes in AfD discussions and the discussions' outcome

As shown in Figure 1, the number of keep and delete votes in an AfD discussion in our dataset ranges from 1 to 23, represented by the column number c and row number r respectively. A cell represents all the AfD discussions that had c keep votes and r delete votes. The cell color reflects the aggregated outcome for that specific cell: the greener/redder the cell is, the more AfD discussions with keep/delete outcome the cell contains. The yellow cell means that it contains equal or close to equal number of keep outcome discussions and delete outcome discussions. Take the green cell in the first row and fifth column as the example. The cell represents this type of AfD discussions: the discussion that has one participant who suggests to delete the article and five participants who suggests to keep the article. The dark green color suggests that most of such discussions kept the article in the end. If one hovers over the cell, the exact number and percentages of keep-outcome and delete-outcome discussions will be displayed on the interface.

Figure 1 shows that a cell's color moves from red to green when we look from the bottom left-hand corner to the top right-hand corner. This indicates that the outcome of an AfD decision is in general consistent with the majority vote rule: the more keep votes than delete votes a discussion has, the more likely its outcome is to keep the article. Nevertheless, it is evident that the difference exists. An election dataset that follows the majority vote rule will have a green upper triangle and a red lower triangle and the diagonal that separates the two will be yellow indicating equal number of opposite votes. This diagram, in contrast, does not follow this pattern. Specifically, the darkness of the color varies in the upper triangle. Similarly, there are orange (lighter red) cells besides red in the lower triangle. In addition, in both triangles there are yellow cells. A striking example is the cell at the fifth row and seventeenth column. It is red, which means that among the AfD discussions that had 17 keep votes and only 5 delete votes, majority of the final decisions were still to delete the article. This illustrates that the decision is not made simply by majority vote in the discussion.

This visualization also offers a bar chart (see Figure 1) that uses the total number of an AfD's votes as the X-axis and the number of discussions as the Y-axis. The red bar represents the AfDs with delete outcomes, and the green ones keep outcomes. It shows that majority of the discussions in this database had five votes in total, and around 80% of these AfDs had delete outcome.

To better understand the distribution of the AfD discussions according to the number of keep and delete votes, we developed another interactive heat map that paints the color of the cell according to the total number of the discussions in the cell (http://www.mandanemedia.com/afd/view/diagram2.php). On the heat map, the lighter the cell's color, the less number of AfD discussions contained in the cell. The number of AfD discussions contained in a cell ranged from 0 to 7019. One can hover over a cell to view the exact amount of AfDs contained in it. The cells in the first column up to row 5 contains about 56% of the AfD. The cells within columns 1-5 and rows 1-6 contain 91% of the AfDs. The cells within the columns 1-10 and the rows 1-10 contain 98% of the AfDs. As the pattern suggests, most of the AfDs had less than 5 keep votes and 5 delete votes.

4.2 The Policies Mentioned in the Participants' Comments

Helping new editors get familiar with the Wikipedia policies in AfD discussions is important [15]. We developed an interactive visualization to offer an overall understanding about the policies

that the participants mentioned during the AfD discussion (http://www.mandanemedia.com/afd/view/diagram3.php). Specifically, we used a sunburst diagram to represent the policies mentioned in the discussion content and their relative frequencies. A sunburst diagram is similar to a tree diagram, except that it uses a radial layout. The root of the tree is in the center of the sunburst diagram, with the leaves on the circumference around it. Our sunburst diagram consists of three tiers: the first tier, which is the innermost circle of the diagram, represents the highest level of the policies: notability vs non-notability policies. The second tier, which is the middle circle of the diagram, represents the middle of the hierarchy; and the third tier, which is the outermost circle of the diagram, represents the bottom of the hierarchy. The length of the arc of each policy corresponds to the percentage of this policy in the participants' comments in the database.

The interaction and concept of this visualization were inspired from the Visual Information Seeking Mantra: overview first, zoom and filter, then details-on-demand [16]. When one hovers over a particular policy on the diagram, the details about the policy's percentage is displayed and the remainder of the diagram fades. The percentage displayed in the center of the sunburst diagram becomes the percentage of this policy mentioned in the comments. If this policy has child policies, this percentage considered the child policies as well. In addition, when one hovers over a policy in the sunburst diagram, the right-hand side of the diagram displays the policy title, the percentage of the AfD discussions that mentioned the policy, and the percentages of the delete, keep, and other comments in the comments that mentioned the policy respectively. In addition to the policy itself, such information of its parent's policy is displayed as well. Figure 2 shows an example of this situation when one hovers above the policy "Wikipedia: CRYSTAL" (in purple).

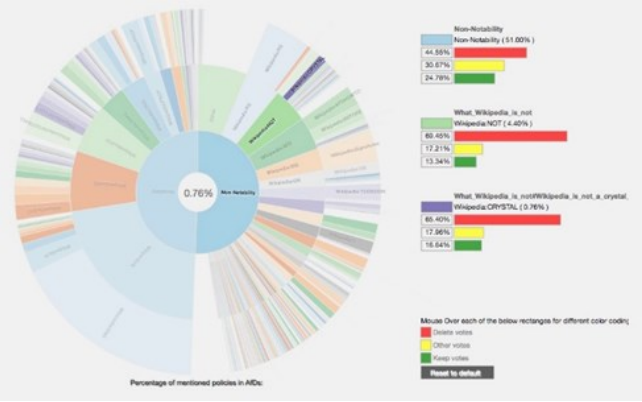


Figure 2: The policies in the AfD votes; the figure also shows the output when one hovering above the particular policy Wikipedia:CRYSTAL

With this visualization, we learned that of the mentioned policies in the discussion data, 49% are notability policies. Also, 59% of the comments that mentioned these policies are for deleting the articles. Among the comments that mentioned non-notability policies, on the other hand, 55% of them are for keeping the article or suggestions other than deleting it. This observation indicates that when the purpose of mentioning notability policies in the comment tends to be to show that the article violates the notability requirement, and the purpose tends to be to "save" the article when non-notability policies are mentioned.

This visualization offers another interaction feature. There are three colored squares on the right-hand lower corner of the diagram representing keep, delete, and other votes. When one clicks on one of them, the diagram visualizes the density of this type of votes in the AfD discussions that mentioned a policy with the darkness of the color. The darker the shade, the higher density of the vote in the discussions contained in the shade. Figure 3 displays the diagram that shows the density of the delete votes in the discussions.

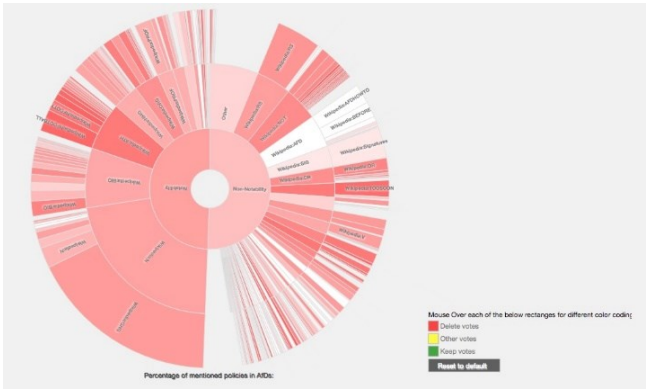
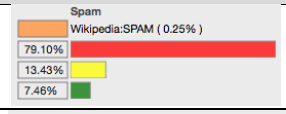
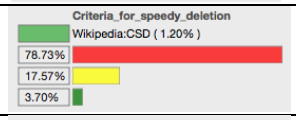
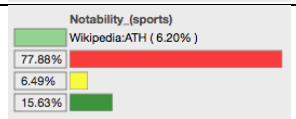
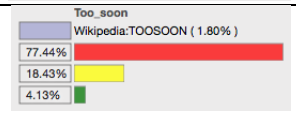


Figure 3: The density of the AfD delete votes that contained the policies

Figure 3 shows that at the second level of the policy structure (i.e., the second tier of the diagram) four policy categories had the darkest color for this type of comment. For each of the four policy categories, the percentage of its delete comments was over 75%. These policies are Wikipedia:SPAM, Wikipedia:CSD, Wikipedia:ATH, and Wikipedia:TOOSOON. Table 1 shows the percentages of the different type of comments (i.e., delete, keep, or other comments) and the percentage of the comments that mentioned one of these four policies in the discussion data. Wikipedia:SPAM and Wikipedia:CSD provide guidelines on what considered to be a spam in Wikipedia and on the criteria of speed deleting an article. The high percentage of delete votes in the comments that mentioned these policies suggest that when these policies were mentioned in the votes, often the participants felt that the articles were spams or should go through the speedy deletion mechanism instead of AfD. Wikipedia:ATH offers criteria on whether an article about an athletic person or a sports player can be included in Wikipedia. The high percentage of delete votes in the comments that mentioned this policy means that when participants mentioned this policy in the deletion discussion often they suggested delete the article. This

implies that when such articles were proposed for deletion, it is often because their authors were unfamiliar with this policy.

Table 1: Percentages of different types of AfD votes that mentioned this policy category

The Policy Category	Percentage of the comments that mentioned this policy category in the comments that mentioned policies	Percentages of keep/delete/other comments in the comments that mentioned this policy category
Wikipedia:SPAM	0.25%	
Wikipedia:CSD	1.2%	
Wikipedia: ATH	6.2%	
Wikipedia:TOOSOON	1.8%	

Sometimes a topic might be notable, but there is no independently reliable source to show it, e.g., an event that just happened and its notability and/or impact have yet to be established. Wikipedia:TOOSOON offers guidelines on these types of situations. The table shows that when this policy was mentioned in one's comment, one's suggestion is often to delete the article. This also implies the author(s)' lack of familiarity with this policy when writing an article that fits into this situation.

At the second level of the policies, three policy categories had over 70% of keep comments in the comments that mentioned them. They are Wikipedia:POINT, Wikipedia:ATATP, and Wikipedia:System bias. They were also rarely mentioned in the AfD discussions, as the percentages of the comments that mentioned them were very low (0.27%, 0.2%, and 0.18% respectively). Interestingly, all three policies are about the discussion norms and/or the community culture/bias. For example, Wikipedia:POINT offers an English Wikipedia behavioral guideline that describes the standard that are accepted by the community and that the editors should attempt to follow. The pages show various example situations as well as what an editor should and should not do. This suggests that when such policies were mentioned in the discussions, the participants' intention was to help "improve and/or save" the article by "citing" these guidelines.

At the third level of the policies, Wikipedia:BEFORE had the highest percentage of keep comments in the comments that mentioned it. Also, about 1.4% of the comments that mentioned a policy mentioned Wikipedia:BEFORE. In these comments, only 3.51% were for delete, 73.54% for keep, and 22.95% for other. This policy describes what one needs to be considered before nominating an article to be deleted (https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion#Before_nominating:_checks_and_alternatives). The high percentage of keep comments implies that in a lot of situations when this policy is mentioned it's because the participant believes the nominator made a mistake in nominating the article. The low percentage of mentioning this policy in the dataset implies that such mistake seldom happens.

Lastly, when one clicks on the title of a policy on the right-hand site information panel, the Wikipedia page for that policy opens in a separate window or tab. This feature is designed to support the scenario when a user wants to check the details about the policy while browsing this visualization.

4.3 The Categories of the AfDs

An earlier study with a small sample suggests that the article's category and the decision of its AfD discussion is correlated [21]. Interested in the distribution of various categories in the proposed articles, we developed another sunburst diagram that shows the categories of the articles proposed for delete (http://www.mandanemedia.com/afd/view/diagram4.php). It has the similar kind of interaction mechanisms as the policy sunburst diagram. For example, one can quickly identify which category of the articles receives the highest/lowest percentages of the keep or delete comments. This tree has four levels with seven categories at the highest level, i.e., the innermost level: Culture and the Arts, General reference, Geography and places, Health and Fitness, Religion and belief systems, Society and Social sciences, and Technology and Sciences.

From this visualization, we learned that three categories have the highest percentages of the delete comments, and they are Martial Art, Football, and Sportspeople. Specifically, 0.98% of the AfDs in our dataset are about Martial Art, and in these discussions, 84.65% of the comments were for delete. Football articles take about 1.5% of the AfDs, and 77.34% of their comments were for delete. About 3.0% of our AfDs were

sportspeople articles, and 77.47% of their comments were for delete. This suggests that when articles of these categories are proposed for Deletion Discussion, likely they will receive a lot of delete comments.

It is also observed that three categories tend to have higher percentage of keep comments: History, Crime, and Architectures. The articles in these categories appear rarely in our AfD dataset, as their percentages in the dataset are 0.57%, 0.57%, and 0.48% respectively. While their keep comments do not take the majority of their votes (37.45%, 35.20%, and 33.48% respectively), they are still higher than other categories. Also, in these three categories, the other comments percentages are moderate (23.27%, 21.66%, 20.68%). When we combine the percentages of keep and other comments together, they are the majority in the comments of these categories. This implies that articles of these categories tend not to be proposed for deletion, and if they are, there are preferences not to delete them.

According to our coding scheme of a participant's comment (see the first paragraph of section 4), a comment may suggest to keep, delete, or offer other suggestions. A common *other* suggestions is to redirect the page to an existing article of the same topic or merge the two articles. Interestingly, the categories that had the highest percentages of other comments all belong to the main category Culture and the Arts. This suggests that articles about Culture and Arts may be more likely to have the situation that the same topic was written twice or more and they had to be merged or redirected, than the other articles.

5 The Mapping of Article Categories and Mentioned Policies

We developed an interactive visualization to offer an overall mapping between the articles' categories and the policies mentioned in their AfD discussions (http://www.mandanemedia.com/afd/view/diagram5.php). This mapping between the categories and the policies was done at the second layer of policy and category hierarchy. Both layers provide an overview of the policy and category respectively, yet enough details are shown as well. When one hovers over the name of a category on the visualization, the policies mentioned in the AfD discussions within this category are highlighted on the right side, as well as the frequency distribution of these policies. For example, Figure 4 shows that 6% our dataset's AfDs are music articles. In these AfD discussions, some comments mentioned policies. Specifically, 8% of the mentioned policies are Notability (people), 31% are Notability, 43% are Notability (music), 10% are Identifying reliable sources, and 8% are Deletion_process. One can open the Wikipedia page about a specific policy by clicking on the name of the policy from this visualization.

6 Discussion

Wikipedia's Article for Deletion (AfD) discussions are critical for maintaining and monitoring the quality of Wikipedia articles. A dataset suitable for computational analysis of the AfD content is much desired given the large amount of unstructured AfD discussion content available on the web. To respond to this research need, we developed an open database that parsed and stored AfDs' discussion content in a relational database for further indexing and processing. Many important but unexplored AfD content analysis can be studied with this open access dataset. For example, [21] suggested that the presence of other votes has influence on the act of saving the articles. But the authors only analyzed 229 AfD discussions. With this dataset, one can explore this relationship at a much larger scale.

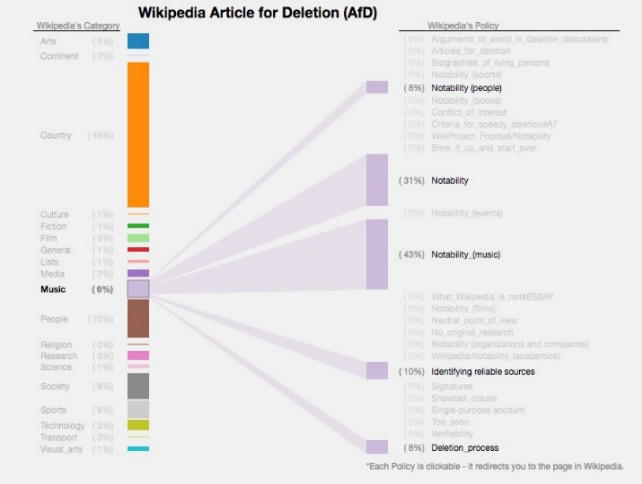


Figure 4: The mapping between the AfDs' categories and the policies mentioned in their discussions; the screenshot shows that music category is hovered above

A Wikipedia article could be nominated multiple times for AfD discussions, and the outcomes can be different from these discussions. In the dataset, about 1000 articles were nominated more than once and went through the AfD discussion process. This presents another interesting aspect to explore - to investigate these articles to understand the rea-sons behind multiple nominations and to compare the different discussions of the same article. We also found that there are AfDs that had delete out-comes but the articles still exist. It is possible that these articles were created again after they were deleted, and they were not proposed for deletion again or not yet. With this dataset, one can explore whether these articles were edited by the same person/people, the reasons of the delete outcomes, and the motivations of writing the same article again after it was deleted by the community.

With this dataset, we developed several interactive visualizations to explore the relationships between the outcomes of the AfD discussions and the number of keep/delete comments in the discussions and to obtain an overview of the policies mentioned in the discussions and the categories of the articles that were proposed for AfD. While some of these aspects were studied in the previous studies (e.g., [21]), we used a much larger dataset (about 40, 000 discussions). The benefits of our interactive visualizations are three folds. First, they present a way to visualize the Wikipedia's community memory in deletion

discussions. Second, these visualizations offer interesting insights on AfD discussions as a research exploration tool for Wikipedia AfD researchers. For example, in the first visualization, the cell in the fifth row and seventeenth column is red, which means that among the AfD discussions that had 17 keep votes and only 5 delete votes, majority of the final decisions were still to delete the article. To find out what happened in these discussions, one can go back to the dataset for further analysis. Third, there is a potential that these visualizations benefit Wikipedia community, especially to the new editors. Previous studies suggest that new editors need tremendous help in realizing the relevant policies and guidelines when editing an article or participating in an article's deletion discussion [6, 14, 15]. In Wikipedia discussions, participants can refer to policies to justify their viewpoints, because policies are supposed to be the shared understandings of how things should be done in the community. The confusions occur among the participants especially the new editors because such common ground is not established with them. By giving an overview and pointers of the mentioned policies in the discussions, new editors can be better aware of the existence of these shared understandings and get acquainted with them more quickly. For example, with the mapping visualization, new author(s) get an understanding of the Wikipedia policies and guidelines they should be familiar with given the type of the current article they are editing. If they need more details about the policies, they can navigate to those Wikipedia pages conveniently from the visualization.

One visualization analysis had an implicit assumption that the policies have stayed the same in the two years of discussions. Wikipedia's policies are the essential parts of AfDs, and the community changes, revises or creates policies in an ongoing basis. So when and how the policies were mentioned in the discussions might be affected by these changes. One future task is to visualize the changes of the policy articles and examine the correlation relationship between the use of the policies and the versions of the policies.

While our AfD dataset offers close to 40, 000 organized AfD discussions, the size is still very small given the population of these discussions. With the database being open access, we encourage researchers to put more AfD discussions in the current dataset and share them in the community.

7 Conclusion

Wikipedia's article for deletion (AfD) discussions offer a communication space for the community members to deliberate about whether or not an article should be deleted from the site. The study of these discussions offers insights on the quality control mechanisms and processes in online peer production systems. To encourage the use of natural language processing and other computational techniques to study the discussion content, we developed an open SQL database with 39,177 discussions. In this paper, we introduced this dataset including the raw data source, the challenges we faced in automatic processing and organizing the data and our strategies to address them. The purpose is to give a comprehensive view of this dataset for those researchers interested in using it in their computational analyses. We call for research studies about the AfD discussions that leverage this open database and suggested several research ideas to explore. As an example, we developed interactive visualizations to gain an overall understanding on the relationship between the deletion outcome and the type of discussion comments, on the Wikipedia policies mentioned in the discussion, and on the categories of articles this dataset include. With these interactive visualizations being open access as well, we hope they become useful resources for Wikipedia community members and for interested researchers.

8 ACKNOWLEDGEMENTS

We gratefully acknowledge the support by the second author's Discovery grant from Natural Science of The Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Arazy, O, Yeo, L., Nov, O.: Stay on the Wikipedia Task: When task - related Disagreements Slip into Personal and Procedural Conflicts. Journal of the Association for Information Science and Technology, 64(8), 1634-1648 (2013)
[2] Biuk-Aghai, R. P., & Cheang, F. H. H. 2011. Wikipedia category visualization using radial layout. In Proc. of the 7th International Symposium on Wikis and Open Collaboration, 193-194
[3] Biuk-Aghai, R. P., Chan, R. C. K., Si, Y. W., & Fong, S. 2013. Visualizing recent changes in Wikipedia. Sci. Ch. Inf. Sci., 56(5), 1-15. 2013
[4] Boldi, P., &Monti, C. 2016. Cleansing Wikipedia Categories using Centrality. In Proc. of the 25th International Conference Companion on World Wide (WWW '16 Companion), 969-974.
[5] Ehmann, K., Large, A., Beheshti, J.: Collaboration in context: Comparing article evolution among subject disciplines in Wikipedia. First Monday, 13(10) (2008)
[6] Geiger, R. S., & Ford, H. 2011. Participation in Wikipedia's article deletion processes. In Proc. of the 7th International Symposium on Wikis and Open Collaboration, 201-202
[7] Iandoli L., Quinto I., Liddo, A. D., & Shum, S. B. 2014. Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. Int. J. Hum. Comput. Stud., 72 (3), 298-319.
[8] Iosub, D., Laniado, D., Castillo, C., Morell, M. F., Kaltenbruner, A.: Emotions under discussion: Gender, status and communication in online collaboration. PloS one, 9(8), e104880 (2014).
[9] Joyce, E., Pike, J., Butler, B.: Keeping eyes on the prize: Officially sanctioned rule breaking in mass collaboration systems. In Int. Proc. of Conference on Computer Supported Cooperative Work (CSCW). 1081–1092, ACM, San Antonio, USA (2013).
[10] Lam, S. T. K., & Riedl, J. 2009. Is Wikipedia growing a longer tail? In Proc. of the ACM 2009 International Conference on Supporting group work, 105-114
[11] Lam, S. K., Karim, J., & Riedl, J. 2010. The effects of group composition on decision quality in a social production community. In Proc. of the 16th ACM International Conference on Supporting Group Work, 55-64
[12] Reagle Jr, J. M.: Do as I do:: authorial leadership in Wikipedia. In: Int. Proc. of Symposium on Wikis, 143-156, ACM, Montreal, Canada (2007)
[13] Rezgui, A., Crowston, K.: Stigmergic coordination in Wikipedia. In: Int. Proc. of OpenSym, (August 22–24, 2018, Paris, France), Retrieved from http://www.opensym.org/wp-content/uploads/2018/07/OpenSym2018_paper_34.pdf (2018)
[14] Schneider, J., Passant, A., & Decker, S. 2012. Deletion discussions in Wikipedia: Decision factors and outcomes. In Proc. of the Eighth Annual International Symposium on Wikis and Open Collaboration, DOI=http://dx.doi.org/10.1145/2462932.2462955
[15] Schneider, J., Samp, K., Passant, A., & Decker, S. 2013. Arguments about deletion: How experience improves the accept-ability of arguments in ad-hoc online task groups. In Proc. of the 2013 conference on Computer supported cooperative work, 1069-1080
[16] Shneiderman, B. 1996. The eyes have it: A task by data type taxonomy for information visualizations. IEEE Symposium on Visual Languages, 336-343
[17] Stefaner, M., Taraborelli, D., Ciampaglia, G. L. 2011. Notabilia - Visualizing Deletion Discussions on Wikipedia. http://notabilia.net
[18] Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. 2008. Information quality work organization in Wikipedia. J Am Soc Inf Sci Technol, 59(6), 983-1001.

[19] Taraborelli, D., & Ciampaglia, G. L. 2010. Beyond notability. Collective deliberation on content inclusion in Wikipedia. In Proc. of the 4th IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO), 122-125

[20] Wattenberg, M., Viégas, F. B., & Hollenbach, K. (2007). Visualizing activity on Wikipedia with Chromograms. In Proc. of the 11th IFIP TC 13 international conference on Human-computer interaction (INTERACT'07), 272-287

[21] Xiao, L., & Askin, N. 2014. What influences online deliberation? A Wikipedia study. J Am Soc Inf Sci Technol, 65(5), 898-910.

[22] Xiao, L., & Nickerson, J. 2019. Imperatives In Past Online Discussions: Another Helpful Source For Community Newcomers? Proceedings of the 52nd Annual Hawaii International Conference on System Sciences (Jan. 8 – 11, Grand Wailea, Maui, USA), IEEE

Inferring Advertiser Sentiment in Online Articles using Wikipedia Footnotes

Shaunak Mishra
Yahoo Research
shaunakm@verizonmedia.com

Aasish Pappu*
Spotify Research
aasishp@spotify.com

Narayan Bhamidipati
Yahoo Research
narayanb@verizonmedia.com

ABSTRACT

Online advertising platforms in partnerships with media companies typically have access to an online user's history of viewed articles. If a concerned brand (advertiser) plans to run advertisement campaigns on users exposed to negative articles, it is essential to first identify articles with negative sentiment about the brand. For an advertising platform, scalable identification of such articles with little human-annotation effort is necessary for launching campaigns soon after an advertiser signs up. In this context, generic sentiment analysis tools suffer from the lack of contextual world knowledge associated with the advertiser. Human annotation of articles for supervised approaches is laborious and painstaking. To address these problems, we propose the use of publicly available Wikipedia footnote references for an advertiser, and propagate their sentiment to several articles related to the advertiser. In particular, our proposed approach has three components: (i) automatically find Wikipedia references which have negative sentiment about an advertiser, (ii) learn distributed representations (doc2vec) of article texts referred in footnotes and other unlabeled articles, and (iii) inferring sentiment in unlabeled articles using label propagation (from references) in the doc2vec space. Our experiments spanning three real brands, and data from a major advertising platform (Yahoo Gemini) show significant lifts in sentiment inference compared to existing baselines. In addition, we share valuable insights on how article sentiment influences the online activities of a user with respect to a brand.

CCS CONCEPTS

• **Information systems** → *Online advertising*.

ACM Reference Format:
Shaunak Mishra, Aasish Pappu, and Narayan Bhamidipati. 2019. Inferring Advertiser Sentiment in Online Articles using Wikipedia Footnotes. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316752

1 INTRODUCTION

A brand's image is as good as its active users perceive it [7]. It is found that active and loyal users of a brand can influence the attitude of former users and yet-to try users. Thus, when a brand suffers negative press, it may consider advertising policies should aim to keep the loyal users in confidence at the same time persuade non-users to convert to users. The challenge could be that a brand may face a public relations crisis, thereby online users are exposed to critical articles thus it impacts brand's revenue [33]. Such articles with negative sentiment about the brand, may influence users who were loyal to the brand in the past, and even nudge them towards the brand's competitors. Among other means to uplift its image [25] and retain its loyal users, the brand may resort to positive advertising. In addition, to be more effective, the brand (advertiser) can focus on the set of users who were exposed to negative sentiments about it [17].

Major advertising platforms typically have collaborations with (or are a part of) online media companies which publish articles, and track users viewing those articles. When an advertiser is interested in finding a segment of users exposed to negative sentiments, it is essential to first identify the articles with negative sentiment on the advertiser.

Criticism [edit]

Major litigation [edit]

Main article: Uber protests and legal actions

Criminal probe due to use of Greyball [edit]

Uber developed an internal software tool called Greyball, which uses data collected from the Uber mobile app oth tool was used starting in 2014. By showing "ghost cars" driven by fake drivers to the targeted individuals in the U cancel rides requested by those individuals, Uber was able to avoid giving rides to known law enforcement office journalism by *The New York Times* and the resulting report, published on March 3, 2017, made public Uber's use city code enforcement officials in Portland, Oregon, Australia, South Korea, and China.[189] At first, in response to deny rides to users who violate Uber's terms of service, including those involved in sting operations.[189][190] Acc view for individual riders, enabling Uber to show that same rider a different version". Uber reportedly used Greyb whether a user frequently opens the app near government offices, a review of social media profiles by Uber empl credit cards associated with the Uber account.[189]

On March 8, 2017, Uber admitted that it had used Greyball to thwart government regulators and pledged to stop

In May 2017, the United States Department of Justice opened a criminal investigation into Uber's use of Greyball operations.[193][194][195][196]

Figure 1: Screenshot of *criticism* section of the Wikipedia page on Uber. The footnotes mentioned in this section are references to articles with negative sentiment on Uber.

Given a large (web scale) collection of online articles on an advertiser, to automatically predict the sentiment of articles, it is preferable to use a method which is: (i) scalable, (ii) domain (advertiser) oriented (iii) afresh with latest topics associated with advertiser, and (iv) distant supervised. For an advertising platform, such preferences are driven by ad-campaign efficiency and low latency while launching campaigns once an advertiser signs up.

In the context of the preferences mentioned above, existing approaches for sentiment analysis are open-domain or trained on generic datasets e.g., Yelp or IMDB. They may be ineffective in a niche domain such as texts about an advertiser. Labeled data

*Work done while the author was at Yahoo Research.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316752

for a niche domain is rare, and expensive to obtain. Most existing approaches for sentiment classification are either supervised models or use transfer learning, hence they are not viable. Furthermore, to keep up with new topics, the annotation task needs to be done on a regular basis on articles with such new topics for every advertiser's domain. Given such shortcomings, we explore a distant-supervised approach that leverages Wikipedia to predict sentiment on unlabeled news articles of an advertiser. A Wikipedia page on an advertiser typically has a section with negative sentiment (*e.g.*, the criticism section), and the footnotes (references to other online articles) mentioned in such a section are human-curated examples of articles with negative sentiment about the advertiser (illustrated in Figure 1. Such footnotes not only provide labeled examples of negative articles about the advertiser, but also keep up with current topics since Wikipedia articles on major entities get updated in a matter of minutes. Footnotes from other (non-criticism) sections in the Wikipedia page on the advertiser can serve as examples of positive (or neutral) articles on the advertiser.

In this paper, we propose an approach which automatically collects the text in Wikipedia footnote articles for an advertiser, and labels them negative or positive depending on the section they are referred in. For the task of inferring sentiment in unlabeled news articles on an advertiser, we first obtain doc2vec [19] embeddings of the footnote articles and the unlabeled articles, and then propagate sentiment from the (labeled) footnote articles to the unlabeled articles. Our main contributions can be summarized as follows:

(1) a scalable sentiment analysis method using Wikipedia footnotes which achieves significant lifts in average precision (AP) and area under ROC curve (AUC) compared to baselines (as high as 19% lift in AP for an advertiser),
(2) an analysis of how the (predicted) sentiment in articles affects online interactions (ad clicks and purchases) of users with the advertiser. We quantitatively validate the intuitive expectation that users exposed to positive sentiments tend to have a higher click through rate (CTR) and conversion rate (CVR) for ads corresponding to the advertiser. In addition, we show that the sentiment for a particular advertiser may or may not carry over to the advertiser's competitors in terms of ad interactions.

2 RELATED WORK

In this section, we cover prior work related to sentiment analysis tools, brand specific sentiment inference, and user behavior models in online advertising.

2.1 Sentiment analysis tools

Sentiment analysis has a huge body of literature, and many state-of-the-art methods are now readily available as tools for text classification; this includes (i) Senticnet [8, 9], (ii) Stanford's CORENLP [2, 31], (iii) NLTK-SentiWordNet [1, 5], (iv) TextBlob [3, 23], (v) VADER [4, 15], and (vi) Polyglot [10]. In this paper, the six tools listed above serve as baselines for comparison against our proposed methods, and additional details regarding each of these methods are described in Section 4.2. Apart from the methods listed above, there has been recent work focusing on review data sets (*i.e.*, IMDB reviews and Amazon reviews) using (supervised) deep learning methods [21, 26, 28, 32, 36]. We do not consider such methods in our paper as we focus on: (i) brand specific sentiment in online news articles, and (ii) unsupervised and semi-supervised methods for inferring sentiment.

2.2 Brand specific sentiment

In terms of inferring sentiment specific to a brand, existing work broadly spans two classes: (i) sentiment prediction using Twittter data (*i.e.*, tweets on a brand) [11, 14, 18, 34], and (ii) investor sentiment towards a brand leading to stock price movement [13, 20, 30]. Specific details for prior work in both the classes are provided below; we first go over prior work using Twitter, followed by work on investor sentiment.

2.2.1 Sentiment analysis using Twitter data. In [14], a supervised approach was proposed for predicting consumer sentiment towards a brand based on tweets. The approach involved constructing a Twitter-specific sentiment lexicon (including tokens related to a brand), which was used to create feature vectors fed to a support vector machine (SVM). In [34], a similar approach was used to extract brand specific features from a smaller set of tweets for mobile companies in Indonesia. The authors measured customer satisfaction for five products for each of the three brands considered in the paper. In [18], 150K tweets were analyzed for linguistic structure and key phrases that convey brand sentiment. The authors also analyzed corporate accounts of specific brands for frequency, timing, and content of their tweets. In [11], tweets were analyzed based on a fixed lexicon of emotional words. The aggregate sentiment of emotional words [27] in a tweet was visualized for corporate users and used to monitor evolving sentiment for brand-specific events.

2.2.2 Investor sentiment and stock prices. Financial news is an invaluable resource to study investors' sentiment towards stocks, and their correlation with stock price movements. Most works described below study the correlation between a public company's evolving sentiment, and their stock price movement. In [30], an active learning approach was introduced to forecast stock price movements via sentiment analysis of stock-related tweets. The authors used an SVM classifier to predict sentiment, and Granger causality test was used to validate stock sentiment as an indicator for its price movements. In [13], a metric based on lexical cohesion was proposed; this measured the sentiment, intensity, and polarity of text. This metric was shown to have strong correlation with human judgments in finance news. In [20], the authors built a stock price prediction framework using the Harvard psychological dictionary and Loughran-McDonald finance sentiment dictionary (for feature engineering). News documents were represented as feature vectors, where the features were tokens in the above lexicons.

In spirit, our focus in this paper is similar to the works described above: inferring brand (advertiser) sentiment in online news articles, and studying their effect on online advertising. The major difference our work introduces is the use of Wikipedia (text in footnote articles in particular), which provides a small set of labeled examples enabling unsupervised and semi-supervised methods for sentiment classification. To the best of our knowledge, there has been no prior work on using Wikipedia footnotes towards brand sentiment classification. In addition, we also study the influence of

the inferred sentiment on users in the context of online advertising (background given below).

2.3 Online advertising and brand sentiment

In a standard online advertising setup, predictive models are employed to predict user behavior (ad clicks and conversions/purchases) when a particular ad is shown to the user; such models are used to select relevant ads as well as for optimizing revenue from ads [6, 24]. In particular, CTR prediction models predict the chances of a click (for a user-ad pair) given available user, ad, and context features. In a similar spirit, conversion models predict the chances of a user converting or purchasing from a particular brand (advertiser) after exposure the brand's ad. In this paper, we explore impact of user sentiment towards a brand (computed as an aggregate of sentiments in relevant news articles read by the user) on the ad click and conversion behavior. In theory, such brand-user sentiment features can be readily consumed by CTR and conversion prediction models used in the advertising industry (which can be as simple as logistic regression [6, 24], or more complex deep neural networks as in [16, 29]). Again, to the best of our knowledge, there is no prior work quantizing the effect of news article sentiments on ad click and conversion behavior with respect to a brand (or its competitors in the same product category).

3 PROPOSED METHOD

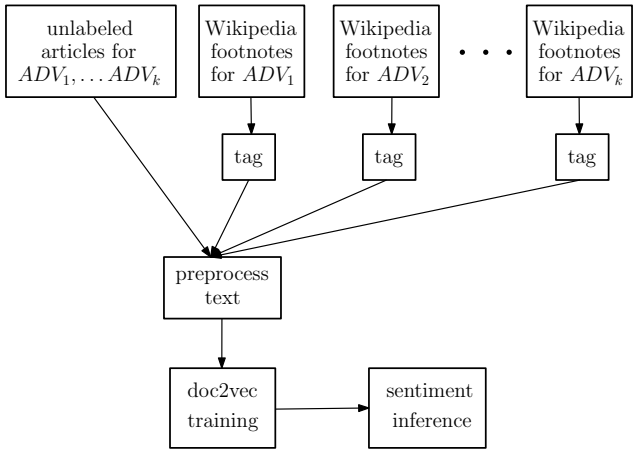


Figure 2: Overview of the proposed architecture.

In this section, we first describe the setup formally in Section 3.1. This is followed by Section 3.2 which gives an overview of the proposed architecture. Section 3.3 covers doc2vec training to obtain embeddings (of labeled footnote articles and unlabeled news articles), and Section 3.4 describes the details of sentiment inference from the doc2vec embeddings.

3.1 Setup

We consider a setup with k advertisers: ADV_1, $ADV_2, \ldots,$ ADV_k. For each advertiser ADV_i there is a corresponding Wikipedia page $Wiki_i$. In $Wiki_i$, $foot_{i,neg}$ is set of footnote articles with negative label (*i.e.*, label = -1) and $foot_{i,pos}$ is the set of articles with positive or neutral label (*i.e.*, label = 1). The labels are assigned based on the section in which the footnotes are mentioned, *i.e.*, only the footnotes in criticism related sections are marked negative[1]. Given such labeled footnote articles (*i.e.*, the entire text in those articles rather than just headlines), we focus on predicting the advertiser sentiment for an unlabeled article $u_{i,j} \in U_i$ where U_i is the set of unlabeled articles for advertiser ADV_i.

3.2 Architecture

Figure 2 shows the proposed architecture. The data collection step includes collecting all Wikipedia footnote articles for all k advertisers. The tag step (labeling) for footnote articles is done on the basis of the section the footnote reference is made (*i.e.*, articles referred to in the *criticism* section of the Wikipedia page are automatically labeled negative; illustrated in Figure 1). The preprocess block cleans the text in the articles (filters out non-English documents, removes punctuation) prior to the doc2vec training step.

The doc2vec block takes as input the *cleaned* text in all labeled footnote articles (across all advertisers) as well as unlabeled articles (details in Section 3.3 below), and produces low dimensional embeddings for each article. Finally using the doc2vec embeddings, sentiment inference is done using the algorithms described in Section 3.4.

3.3 Doc2vec training

We obtain low dimensional embeddings of all articles (labeled footnotes and unlabeled articles) across all advertisers using doc2vec [19]. Doc2Vec learns low dimensional embeddings for words and documents (articles) from a large corpus in an unsupervised manner. In particular, doc2vec embeddings of articles with high semantic similarity tend to be *close* to each other in terms of cosine similarity between the learnt embeddings. In our experiments, we use the vanilla version of doc2vec [19], as well as an enhanced version as proposed below.

3.3.1 Opinion observer enhanced doc2vec (OO+). We use the vocabulary in the seminal *Opinion Observer* work [22] to obtain a set of negative words (denoted by set N_{OO}), and enhance doc2vec training in the following manner. For each negative sentiment footnote article, we identify words/phrases which are also present in N_{OO}, and replace them in the article text by *criticism* tag *e.g.*, *pricey* is replaced by *criticism_pricey*. By doing so, we establish a stronger connection across all negative footnote articles prior to the doc2vec training step. In the remainder of this paper, we will refer to this as the OO+ version of doc2vec (as opposed to the vanilla version described above).

3.4 Sentiment inference

Using the doc2vec embeddings of labeled footnote articles and unlabeled articles on advertisers, we propagate the sentiment from labeled articles (referred in footnotes under *criticism* section) using the two approaches described below.

[1]There is a small degree of label noise in the footnote labeling procedure based on Wikipedia sections; there are cases when a footnote mentioned in a non-criticism section is actually of negative sentiment. In this paper, we do not consider such label noise, but evaluate our proposed methods on an independently annotated data set.

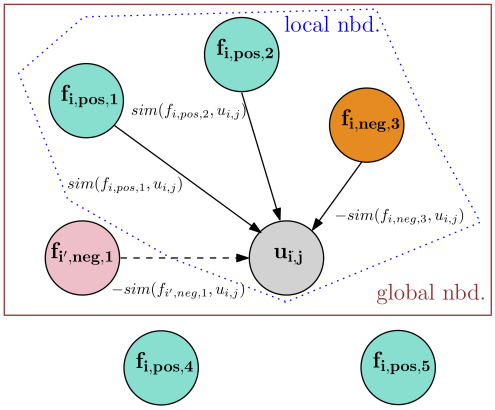


Figure 3: Nearest neighbour sentiment inference in the doc2vec embedding space. In the local mode, the average of the footnote labels of the same advertiser (weighted by cosine similarity) is considered, while in the global mode footnotes across all advertisers are considered.

3.4.1 Nearest neighbour. In the nearest neighbour approach, we simply obtain the average sentiment score of an unlabeled article by a weighted average of the labels in its neighbourhood. The cosine similarity between the doc2vec embedding of the unlabeled article and a labeled article in its neighbourhood is used as the weight during the averaging process. The nearest neighbour approach has multiple parameters which can be used for tuning the approach for each advertiser.

(1) We consider only the top m_i neighbours (ranked by cosine similarity) while doing the weighted average for each unlabeled article for ADV_i.
(2) For an unlabeled article on ADV_i, we can either consider only ADV_i related footnotes (*i.e.*, local mode) or consider all footnotes across multiple advertisers (*i.e.*, global mode) as illustrated in Figure 3.

Intuitively, the global mode may be more effective when the number of footnotes for an advertiser is low.

3.4.2 Label propagation. Label propagation is a family of semi-supervised algorithms based on graph representations. Both labeled and unlabeled instances in the data are considered as vertices, and a (vertex) similarity function is used to assign edge weights between a pair of vertices. At a high level, label propagation algorithms exploit the manifold structure in unlabeled data, and assign (propagate) labels from labeled vertices to unlabeled vertices. In the past, several text-based problems have been formulated in this setting [12, 35, 35]. In this paper, we use the label propagation method proposed in [37], and construct an affinity matrix of document representations (doc2vec vectors) for both labeled and unlabeled news articles in our data set. In particular, we use cosine similarity between the doc2vec vectors as the edge weight between a pair of vertices. The label propagation algorithm iteratively assigns labels in high-density areas of unlabeled documents in the doc2vec space; it also learns parameters using minimum spanning tree heuristic, and entropy minimization to fine-tune label assignment. In a spirit similar to global and local modes in the nearest neighbour approach, we use the term global mode for label propagation when we consider a graph with unlabeled articles and footnote articles across all advertisers. In the local mode, we consider a graph with unlabeled articles and footnote articles only for the concerned advertiser.

adv	# wiki notes	# +ve wiki	# articles	# labeled articles	# +ve labeled
ADV_1	221	125	6921	284	83
ADV_2	82	48	1479	205	150
ADV_3	40	27	4314	257	168

Table 1: Articles data for each advertiser.

4 RESULTS

In this section, we describe results on the accuracy of the proposed sentiment analysis methods, as well as associated user behaviour insights. We first describe our data sources in Section 4.1, followed by Section 4.2 on baseline methods, and Section 4.3 on their performance in comparison to the proposed methods. This is followed by Section 4.4 on insights obtained from user interactions with ads after reading articles spanning diverse sentiments on various brands (advertisers).

4.1 Data

We conducted experiments on a mix of public (Wikipedia) and proprietary data for three advertisers (anonymized as ADV_1, ADV_2, ADV_3). In particular, ADV_1 is an employer for drivers in the transportation domain, ADV_2 is an e-commerce portal, and ADV_3 is a wireless (phone) service provider. The data was collected as described below, and the relevant counts are summarized in Table 1.

4.1.1 Wikipedia data. We collected data on ADV_1, ADV_2, and ADV_3 from Wikipedia (*i.e.*, the text from online articles listed as footnotes in a brand's Wikipedia page). A footnote article associated with the criticism section of a brand's Wikipedia page was marked as having negative sentiment, and the rest were marked positive. The number of footnote articles per advertiser is shown in Table 1 along with the count of footnotes marked positive (or neutral). In addition, a visualization of the doc2vec embeddings of the footnote articles reveals a sense of separability between positive and negative footnote articles as shown in Figure 4.

4.1.2 Data from Yahoo! articles. We collected online articles related to ADV_1, ADV_2, and ADV_3 which appeared in Yahoo Finance, Yahoo Sports, and Yahoo News during the period September 2017 - October 2018. From this collection of online articles, an editorial team selected articles (with higher counts of users who read them) and annotated them with sentiments. The selected articles were primarily about one of the three advertisers in consideration. The count of annotated (labeled) articles per advertiser is summarized in Table 1. For example, for ADV_2 there were 1479 articles, and out of those only 205 were labeled. In addition, we obtained anonymized data from a major advertising platform (Yahoo Gemini) regarding

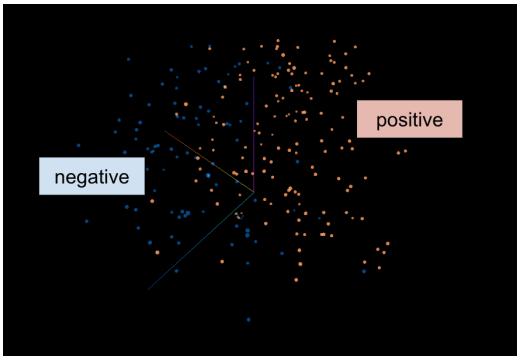


Figure 4: Visualization of 100 dimensional doc2vec embeddings for ADV_1 footnote articles using TensorFlow embedding projector. The blue dots represent footnote articles with negative sentiment, and the orange dots represent footnote articles with positive sentiment.

trails of online user activities, *e.g.*, anonymized history of a user's article views, ad clicks and conversions (*e.g.*, purchases, sign-ups, installs).

4.2 Baseline sentiment analysis methods

To benchmark our results against existing state-of-the-art sentiment methods, we ran six sentiment classifiers on our testing data. We choose four lexicon or knowledge-base based methods and two statistical methods as our baselines.

4.2.1 Senticnet. Senticnet [8, 9] is a semi-automatically constructed sentiment resource using semantic web techniques to glean opinions from natural language. It adopts an energy-based formalism as used in COGBASE to connect multi-word expressions with semantic concepts. Previous versions of Senticnet used an ensemble of graph-mining, and dimensionality reduction algorithms.

4.2.2 Stanford NLP. Stanford's CORENLP provides a sentiment analysis classifier that is based on compositional model over trees using deep learning [2, 31]. The classifier assigns a sentiment score to a sentence in a document where every sentence is represented as a binary tree.

4.2.3 NLTK-SentiWordNet. NLTK-SentiWordNet [1, 5] extends WordNet with numerical annotations for every synset in a collection of 147,306 synsets. These numerical annotations correspond to positive, negative and neutral classes.

4.2.4 TextBlob. TextBlob is a text processing toolkit that includes sentiment classifier as one of its many text analysis tools [3, 23]. It uses a semi-automatically constructed subjectivity lexicon of adjectives, where each adjective has a polarity score (negative is -1.0, positive is +1.0) and a subjectivity score (objective: 0.0 to subjective 1.0). These scores are further annotated for reliability of the assignment (1.0 for human vs 0.7 for automatic).

4.2.5 Vader Sentiment. VADER is Valence Aware Dictionary for Sentiment Reasoning tool. It is a rule-based method that empirically bootstraps a lexicon of tokens [4, 15]. The rules are primarily

advertiser	method	avg precision	AUC
ADV_1	TextBlob	0.295	0.559
ADV_1	VADER	**0.577**	**0.784**
ADV_1	SentiWordNet	0.446	0.692
ADV_1	SenticNet	0.398	0.623
ADV_1	Polyglot	0.462	0.702
ADV_1	Stanford	0.357	0.623
ADV_2	TextBlob	**0.855**	**0.734**
ADV_2	VADER	0.832	0.690
ADV_2	SentiWordNet	0.813	0.633
ADV_2	SenticNet	0.784	0.589
ADV_2	Polyglot	0.832	0.678
ADV_2	Stanford	0.853	0.689
ADV_3	TextBlob	0.748	0.617
ADV_3	VADER	0.798	0.661
ADV_3	SentiWordNet	0.776	0.665
ADV_3	SenticNet	0.721	0.584
ADV_3	Polyglot	**0.803**	**0.677**
ADV_3	Stanford	0.697	0.579

Table 2: Baseline results for each advertiser.

grammatical and syntactical in nature and are aimed to extract intensity of words that convey sentiment. The authors of VADER report state-of-the-art results that match [31] results on various benchmark data sets include Amazon reviews corpus and IMDB review data set.

4.2.6 Polyglot. Polyglot is a multilingual natural language toolkit that includes sentiment classification as one of the many tools[10]. The classifier is based on graph-propagation that connects most frequently used words across 136 languages on Wikipedia. The links/connections between the words help propagate the sentiment across various languages using graph propagation and label propagation methods.

4.3 Evaluation on labeled data

4.3.1 Baselines: Table 2 shows the performance of the six baselines (listed in Section 4.2) on the labeled data for each advertiser. In particular, to obtain the sentiment *score* of each article, we averaged the sentiment score (given per line by the baseline method) across all lines in the article. The performance is measured in terms of average precision and area under the ROC curve (AUC). As shown in Table 2, VADER, TextBlob and Polyglot are the best baseline methods for for ADV_1, ADV_2 and ADV_3 respectively.

4.3.2 Proposed approaches. To compare with the best baseline results for each advertiser, Table 3 shows the evaluation results for the proposed methods (the first row for each advertiser corresponds to the best baseline for the advertiser). As shown in Table 3, for ADV_1, the label propagation method with vanilla doc2vec embeddings, and propagation using only ADV_1 footnote articles (*i.e.*, local mode) has

advertiser	method	doc2vec version	neighborhood	avg precision	AUC
ADV_1	VADER	-	-	0.577	0.784
ADV_1	label prop	OO+	global	0.596	0.844
ADV_1	label prop	OO+	local	0.642	0.865
ADV_1	label prop	vanilla	global	0.618	0.851
ADV_1	label prop	vanilla	local	**0.685**	**0.876**
ADV_1	nearest neighbor	OO+	global	0.613	0.830
ADV_1	nearest neighbor	OO+	local	0.616	0.834
ADV_1	nearest neighbor	vanilla	global	0.614	0.825
ADV_1	nearest neighbor	vanilla	local	0.628	0.823
ADV_2	TextBlob	-	-	**0.855**	**0.734**
ADV_2	label prop	OO+	global	0.800	0.599
ADV_2	label prop	OO+	local	0.779	0.565
ADV_2	label prop	vanilla	global	0.820	0.622
ADV_2	label prop	vanilla	local	0.758	0.508
ADV_2	nearest neighbor	OO+	global	0.783	0.586
ADV_2	nearest neighbor	OO+	local	0.793	0.585
ADV_2	nearest neighbor	vanilla	global	0.797	0.595
ADV_2	nearest neighbor	vanilla	local	0.799	0.586
ADV_3	Polyglot	-	-	0.803	0.677
ADV_3	label prop	OO+	global	0.862	0.803
ADV_3	label prop	OO+	local	0.768	0.668
ADV_3	label prop	vanilla	global	0.851	0.783
ADV_3	label prop	vanilla	local	0.715	0.613
ADV_3	nearest neighbor	OO+	global	**0.878**	**0.808**
ADV_3	nearest neighbor	OO+	local	0.720	0.624
ADV_3	nearest neighbor	vanilla	global	0.861	0.792
ADV_3	nearest neighbor	vanilla	local	0.702	0.596

Table 3: Performance of the proposed methods for each advertiser.

the best performance. A plausible reason behind its success could be the presence of a large number of available footnote articles (*i.e.*, 221) for ADV_1. In contrast, for ADV_3, which has a very few footnote articles (*i.e.*, 40), the nearest neighbor classifier using global mode (*i.e.*, using footnotes for all the three advertisers) has the best performance. In the case of ADV_2, none of the proposed approaches are better than the best baseline (TextBlob); however, label propagation in the global mode comes closest in performance to TextBlob[2]. The results also indicate that the nearest neighbour approaches are usually competitive in performance to the label propagation approaches. This is encouraging on the scalability front since there are very efficient methods (*e.g.*, locality sensitive hashing) for looking up neighboring footnote articles in the doc2vec space.

4.4 User Behaviour Insights

On grounds of scalability, we used the best nearest neighbor approach for each of the three advertisers, and produced sentiment scores for each article associated with an advertiser (*i.e.*, 6921 articles for ADV_1, 1479 for ADV_2, and 4314 for ADV_3). For each advertiser, we identified online users (using data from the advertising platform) who had read online articles about the advertiser in the September 2017 - October 2018 time window (*i.e.*, exposed users). We also identified if the set of exposed users had (i) clicked on the advertiser's ads, (ii) clicked on the advertiser's competitor's ads, and (iii) converted on the advertiser (e.g., purchase from the advertiser). Using the above data, we inferred the (advertiser specific) user sentiment prior to a target event (*i.e.*, advertiser ad click or advertiser conversion or competitor ad click) in the following manner:

$$\texttt{sentiment}_{user,ADV} = \frac{\sum_{\text{article} \in \text{history}} \texttt{sentiment}_{article}}{|\text{articles} \in \text{history}|},$$

where *history* is the set of articles on *ADV* read by the user before the target event, and $sentiment_{article}$ is the sentiment score of the article as obtained by the nearest neighbour approach. If the user did not do the target event in the September 2017 - October 2018 time window, the history includes all articles read by the user on the advertiser. Using the above definition of user sentiment for each advertiser, we obtained the user sentiment across all exposed users and divided them into two groups: positive users and negative users (by using a threshold on the user sentiment score). Table 4 shows the target event rate for positive and negative users (rate normalized by the average target event rate for the entire set of exposed users); the target events in the table include: (i) conversion (*i.e.*, if the user purchased/converted after exposure), (ii) ad click (*i.e.*, if the user clicked on the advertiser's ad after exposure), and (iii) competitor ad click (*i.e.*, if the user clicked on a chosen competitor's

[2]On further examination, we found that the labeled footnotes for ADV_2 had a small degree of label noise, which could be causing inferior performance using our proposed method.

adv.	# exposed in mil.	target event	target event rate	+ve user target rate lift	-ve user target rate lift
ADV_1	2.17	conversion	0.0089	1.0637	0.8749
ADV_1	2.17	ad click competitor	0.1081	1.0400	0.9582
ADV_1	2.17	ad click	0.0676	0.9748	1.0261
ADV_2	0.97	conversion	0.0836	1.0228	0.8817
ADV_2	0.97	ad click competitor	0.1020	1.0274	0.8737
ADV_2	0.97	ad click	0.4821	1.0116	0.9444
ADV_3	1.96	conversion	0.0369	1.0903	0.8915
ADV_3	1.96	ad click competitor	0.1098	1.0116	0.9862
ADV_3	1.96	ad click	0.0258	0.9959	1.0025

Table 4: Impact of user sentiment on ad clicks and conversions.

ad after exposure to the advertiser's articles). For example, as shown in Table 4, there were about 0.97 million exposed users for ADV_2, and out of them the positive users had a normalized conversion rate of 1.0228 while the negative users had a normalized conversion rate of 0.8817. Based on this, the positive users were 16% more likely to convert on ADV_2 compared to negative users. The patterns for ad clicks and conversions with regards to user sentiment are similar across all advertisers; however, when it comes to behaviour towards ads of competitors, there is a difference. For ADV_2, users with positive sentiment are more likely to click on ads of the competitor (*i.e.*, another e-commerce company) than negative users. However, for ADV_3, negative users for ADV_3 are more likely to click on a competitor's ads (*i.e.*, another wireless service provider). These insights are valuable towards consuming sentiment signals in click and conversion models, and serve as useful inputs the advertiser as well.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we explore the use of Wikipedia footnote articles towards advertiser specific sentiment analysis. Using a small number of foot note articles ($\sim$ 200), our approach was able to outperform six competitive baselines for two out of three advertisers in our experiments. The ability to propagate sentiments to a large number of unlabeled news articles on an advertiser, enables us to not only identify users exposed to negative sentiments, but also quantify the impact on ad clicks and conversions. However, as discussed below, a few challenging topics remain open and are directions for future work.

(1) During the data validation phase (via random checks on the foot note labels), we identified Wikipedia footnotes from the *criticism* section which were neutral. We also noticed articles in the non-criticism section which were of negative sentiment. Although small in number, such inconsistencies introduce label noise in our setup. Quantifying such label noise across advertisers (with minimum human involvement), and making our sentiment propagation approach more robust are directions for future work.

(2) In our experiments, we found that for an advertiser with negligible number of footnote articles, it is better to propagate sentiments from footnotes across multiple advertisers (*i.e.*, the global mode in the nearest neighbour approach). Refining this along the lines of transfer learning is another direction for future work.

REFERENCES

[1] NLTK SentiWordNet. http://www.nltk.org/howto/sentiwordnet.html.
[2] Stanford Core NLP. https://github.com/dasmith/stanford-corenlp-python.
[3] TextBlob. https://github.com/sloria/textblob.
[4] Vader sentiment. https://github.com/cjhutto/vaderSentiment.
[5] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.
[6] N. Bhamidipati, R. Kant, S. Mishra, and M. Zhu. A large scale prediction engine for app install clicks and conversions. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM 2017)*.
[7] M. Bird, C. Channon, and A. S. Ehrenberg. Brand image and brand usage. *Journal of Marketing Research*, pages 307–314, 1970.
[8] E. Cambria, D. Olsher, and D. Rajagopal. Senticnet 3: a common and commonsense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.
[9] E. Cambria, S. Poria, D. Hazarika, and K. Kwok. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *AAAI*. AAAI Press, 2018.
[10] Y. Chen and S. Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 383–389, 2014.
[11] E. Colleoni, A. Arvidsson, L. K. Hansen, and A. Marchesini. Measuring corporate reputation using sentiment analysis. In *Proceedings of the 15th International Conference on Corporate Reputation: Navigating the Reputation Economy, New Orleans, USA*, 2011.
[12] B. Dalvi, E. Minkov, P. P. Talukdar, and W. W. Cohen. Automatic gloss finding for a knowledge base using ontological constraints. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 369–378. ACM, 2015.
[13] A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 984–991, 2007.
[14] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.
[15] C. H. E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.
[16] D. Gligorijevic, J. Gligorijevic, A. Raghuveer, M. Grbovic, and Z. Obradovic. Modeling mobile user actions for purchase recommendation using deep memory networks. SIGIR, 2018.
[17] S. J. Hoch and J. Deighton. Managing what consumers learn from experience. *The Journal of Marketing*, pages 1–20, 1989.
[18] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.
[19] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, 2014.
[20] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23, 2014.
[21] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
[22] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05.
[23] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey, et al. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 2014.
[24] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: a view from the trenches. KDD 2013.
[25] S. J. Milberg, C. Whan Park, and M. S. McCarthy. Managing negative feedback effects associated with brand extensions: The impact of alternative branding

strategies. *Journal of Consumer Psychology*, 6(2):119–140, 1997.

[26] T. Munkhdalai and H. Yu. Neural semantic encoders. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 397. NIH Public Access, 2017.

[27] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[28] A. Radford, R. Jozefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

[29] Y. Shan, T. R. Hoens, J. Jiao, H. Wang, D. Yu, and J. C. Mao. Deep crossing: Web-scale modeling without manually crafted combinatorial features. KDD 2016.

[30] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285:181–203, 2014.

[31] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[32] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[33] C. C. Verschoor. Uber culture causes big losses: harassment and mismanagement have led to steep losses for this high-flying company. *Strategic Finance*, 99(3):23–25, 2017.

[34] N. A. Vidya, M. I. Fanany, and I. Budi. Twitter sentiment to analyze net brand reputation of mobile phone providers. *Procedia Computer Science*, 72:519–526, 2015.

[35] M. Yu, S. Wang, C. Zhu, and T. Zhao. Semi-supervised learning for word sense disambiguation using parallel corpora. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 3, pages 1490–1494. IEEE, 2011.

[36] Z.-H. Zhou and J. Feng. Deep forest: Towards an alternative to deep neural networks. *arXiv preprint arXiv:1702.08835*, 2017.

[37] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation.

Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata

Charlotte Rudnik
LIMSI, CNRS
Orsay, France
charlotte.rudnik@limsi.fr

Thibault Ehrhart
EURECOM
Sophia Antipolis, France
thibault.ehrhart@eurecom.fr

Olivier Ferret
CEA, LIST,
Gif-sur-Yvette, F-91191, France
olivier.ferret@cea.fr

Denis Teyssou
Agence France-Presse (AFP)
Paris, France
denis.teyssou@afp.com

Raphael Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

Xavier Tannier
Sorbonne Université, Inserm, LIMICS
Paris, France
xavier.tannier@sorbonne-universite.fr

ABSTRACT

News agencies produce thousands of multimedia stories describing events happening in the world that are either scheduled such as sports competitions, political summits and elections, or breaking events such as military conflicts, terrorist attacks, natural disasters, etc. When writing up those stories, journalists refer to contextual background and to compare with past similar events. However, searching for precise facts described in stories is hard. In this paper, we propose a general method that leverages the Wikidata knowledge base to produce semantic annotations of news articles. Next, we describe a semantic search engine that supports both keyword based search in news articles and structured data search providing filters for properties belonging to specific event schemas that are automatically inferred.

ACM Reference Format:
Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphael Troncy, and Xavier Tannier. 2019. Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316761

1 INTRODUCTION

Information and communication technologies have provided tools and methods to make the production of information more democratic. In this context, journalists and technologists have developed the notion of "data journalism", which takes advantage of structured and numerical data in the production and distribution of news. It also takes advantage of the growing popularity of Linked and Open Data and the development of structured knowledge bases such as DBpedia [5], YAGO [3] or Wikidata [17] to facilitate information analysis and to access a variety of points of view. However, knowledge is still far from being entirely represented in structured databases, and the most prominent way to convey information to the end user is still the free text, complemented by multimedia content.

Intertwining structured and unstructured data in information systems is still an open research problem. In this paper, we present a system for aggregating unstructured news articles and structured data describing events leveraging on the Wikidata knowledge base. This approach makes use of several Information Retrieval and Information Extraction tasks. In the context of Information Extraction, the knowledge associated with news articles can typically be used for training event extractors in a distant supervision mode [12]. From the Information Retrieval perspective, the approach makes it possible to retrieve news articles describing events using either keyword-based queries or filters that typically make use of properties available in knowledge bases. It also allows to query Wikidata and then to read an entire annotated article describing the corresponding event. We implemented a system which is available at http://asrael.eurecom.fr/ and covers the last two tasks.

Figure 1 illustrates the architecture of our system. Events described in news articles are mapped to events from Wikidata (Section 4.1), and attributes from the Wikidata instances are used to annotate the news articles when possible (Section 4.3). Wikidata events belong to specific classes, but these classes are too fine-grained for being used in a search engine. Furthermore, many event classes actually share a similar structure (*i.e.* sets of attributes). For example, a general *election* schema is relevant for describing any type of elections regardless of the more specific Wikidata types such as "*Bundestag election*" (`Q1007356`), "*direct election*" (`Q1196727`), "*Elections in Saudi Arabia*" (`Q4119635`)…. For these reasons, we add a hierarchical clustering step (Section 4.2) to automatically create coarser grained schemas. Finally, we implemented an event-oriented knowledge graph and a search engine able to query and navigate through both the knowledge base and the news articles (Section 4.4).

2 RELATED WORK

Contrarily to WikiNews[1], Wikipedia does not aim to be a news service. However, Wikipedia's Current Events portal (WCEP[2]) provides a set of pages where primarily events but also trends and developments are listed on a daily basis with links to reference

[1] https://www.wikinews.org/
[2] https://en.wikipedia.org/wiki/Portal:Current_events

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316761

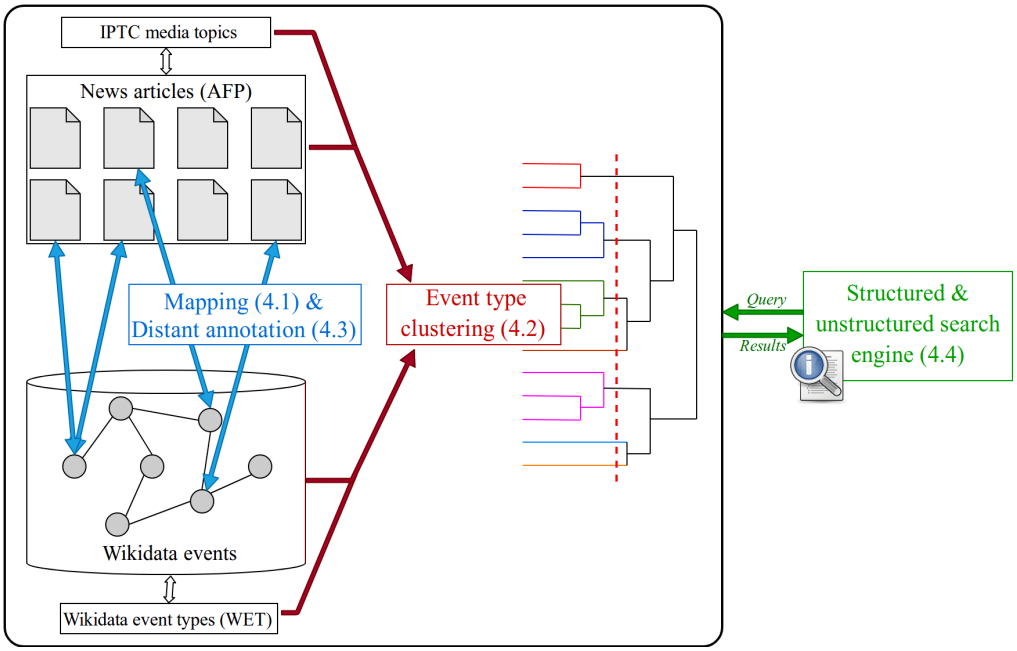


Figure 1: System overview for annotating news articles and enabling structured search.

articles. The WikiTimes project[3] is the first attempt to build a structured and rich knowledge base of news events by harvesting the efforts of the Wikipedia crowd in maintaining WCEP [14]. The WikiTimes knowledge base is represented in RDF and it contains very short descriptions of events that can be filtered by entities, location and time.

Based on this experience, Gottschalk and Demidova have developed the EventsKG knowledge graph[4], a multilingual resource incorporating event-centric information extracted from Wikidata, DBpedia and YAGO, as well as less structured sources such as the Wikipedia Current Events Portal and Wikipedia event lists in five languages [2]. EventsKG re-uses the Simple Event Model (SEM) ontology [16] to describe nearly 700,000 events. However, temporal information is available for 76% of those events and location information for only 12% of them. While many entities are mentioned, extracted and disambiguated in the short descriptions of those events, this is far from being complete. Events are generally weakly categorized and both categorical and numerical data representing the events attributes are rarely extracted. Annotating semantically newsfeeds at scale is being continuously proposed in [13] which maintains the Newsfeed[5] service. Annotations are, however, restricted to named entities that can be extracted by the Enrycher tool[6].

For easing the exchange of news, the International Press Telecommunication Council (IPTC) has developed the NewsML Architecture (NAR), specialized into a number of languages such as NewsML G2 and EventsML G2. As part of this architecture, specific controlled vocabularies, such as the IPTC Media Topics or News Codes, are used to categorize news items together with other industry-standard thesauri. In previous work, we designed an OWL ontology for the IPTC News Architecture and we converted the IPTC NewsCodes into a SKOS thesaurus [15]. IPTC is now publishing itself the IPTC Media Topics in SKOS and has further developed the rNews vocabulary, largely based on Schema.org, for describing news articles. In this work, we re-use the rNews vocabulary to describe the original metadata attached to news articles. Furthermore, we annotate the news articles using properties and entities from Wikidata once events reported in the news have been mapped to existing Wikidata events.

While mapping text to knowledge has been the subject of a large body of work, represented in the recent ages with work such as [7] or all the work about entity linking [1], mapping text to event representations and more particularly news articles to event representations has not been the focus of lots of studies. One exception is [9], followed by [10], which tackles this kind of mapping according to an Information Retrieval perspective through two tasks based on the notion of *Wiki-excerpt*. A *Wiki-excerpt* corresponds to a description of an event built from Wikipedia and contains both a textual description and factual information about the event such as temporal expressions, geolocation and named entities. The first task, *Wiki2News*, starts from a *Wiki-excerpt* and aims at retrieving a set of past news articles about the considered event while *News2Wiki* is the reverse task, consisting in retrieving *Wiki-excerpts* from a set of news articles. The work focuses more specifically on the *Wiki2News* task by designing time-aware language models for supporting the retrieval of past news articles. More recently, the *Wiki2News* task of [9] has been considered under the perspective of the enrichment

[3]http://wikitimes.l3s.de/
[4]http://eventkg.l3s.uni-hannover.de/
[5]http://newsfeed.ijs.si/
[6]http://enrycher.ijs.si/

of Wikipedia from a stream of news by [6]. This work first builds a temporal event chain of the news articles related to the target event and then selects a subset of them according to various representativeness criteria exploited in a learning-to-rank framework. While all the work we have mentioned was based on Wikipedia, our work tackles the *Wiki2News* task by relying on Wikidata as a knowledge base, with a much simpler, still effective, approach.

3 DATA

3.1 AFP News Articles

In this work, we make use of a very large corpus of text newswire written in English provided by the French news agency AFP. More precisely, we use over 2 million articles covering the period 2004-2017. The topics are worldwide news ranging from politics, diplomacy, sports to natural disasters or economy and business. Each document is an XML file compliant with the NewsML standard, containing a title, a document creation time (DCT), a dateline where the article was written, one or several IPTC Media Topics and a set of keywords (slugs), as well as a textual content split into paragraphs.

The main topic of an article is generally a specific event, and sometimes, other older events are referred to in order to look at the current one from a wider perspective. This is why we consider that it is possible to associate an AFP article with one single event. Furthermore, we assume that the title and the first paragraph (lead) describe the event associated with the document. This is a realistic hypothesis since the basic rules of journalism impose that the first sentence should summarize the event by informing on the "5 Ws" (*What*, *Who*, *When*, *Where*, *Why*).

3.2 Wikidata Occurrences

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines [17]. It acts as a central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, WikiNews, and others. In our work, we focus on newsworthy events, *i.e.* something that happens within a locality and a temporality, and that could be described in one or more news articles. In regard to this definition, there are many general and specific classes in Wikidata that are related to events, but all these classes have in common the same parent class named "Occurrence" (`Q1190554`). For example, the event "Cargolux Flight 7933" (`Q3107014`) is an *instance_of* (`P31`) an "aviation accident" (`Q744913`) which is a *subclass_of* (`P279`) an "aviation occurrence" (`Q15733640`) which is a *subclass_of* (`P279`) "occurrence" (`Q1190554`).

In the remainder of this paper, we call Wikidata Event Type (or WET) the value of the property *instance_of* (`P31`) of a Wikidata event. In the previous example, "aviation accident" (`Q744913`) is a WET. An instance can have several Wikidata Event Types.

We lined up on the temporal coverage of the AFP articles corpus and we considered all Wikidata event instances during the period from 2004 to 2017. The event date can be represented by three properties in Wikidata:

- `P585`:*point_in_time* if the event is a one-off event;
- `P580`:*start_time* and `P582`:*end_time* if the event has a duration.

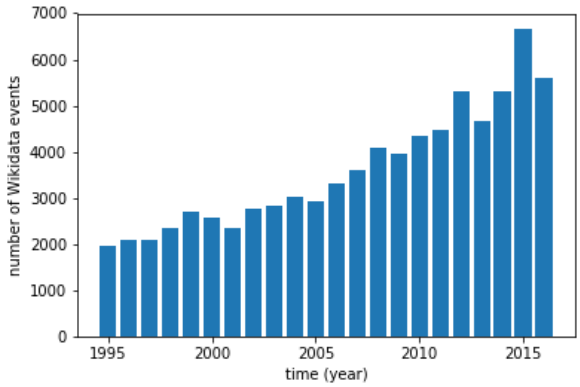


Figure 2: Evolution of the number of Wikidata events over time.

In the second case, we consider the events that have started and ended during the period 2004-2017.

The selection of Wikidata events consists of 60k Wikidata instances. As shown in Figure 2, the distribution over the years is not uniform. The increase in the last few years is explained by a better quality of the data and particularly by the presence of a date in the description of the events, but should not be interpreted as an increase of the number of events happening in the world.

3.3 IPTC Media Topics

The International Press Telecommunications Council (IPTC) maintains a taxonomy of Media Topics[7], which can be seen as a controlled and hierarchical set of indexing keywords. Each article written by AFP is associated with at least one IPTC code by its author. IPTC Media Topics (later called IMTs) give information about the topic of the article and are often linked with a type of event (earthquake, election, crash…), but not always (politics, theatre…). Figure 3 illustrates a part of the IPTC Media Topic hierarchy.

4 APPROACH

Following the description of the data in Section 3, we define that "AFP article" stands for the main event associated with the AFP article and described in the lead, while "Wikidata event" stands for the structured events described in our selection of Wikidata instances. The term 'article' will refer to the news article whereas the term instance will refer to a Wikidata instance.

4.1 Mapping AFP with Wikidata

In order to map Wikidata events to AFP articles, we consider that two mentions of the same event share the following characteristics: same time, same place and same type or category (election, natural disaster, etc.), which we defined as a content similarity.

4.1.1 Scoring function. For mapping a Wikidata instance to an AFP article, we define the following criteria:

[7] http://cv.iptc.org/newscodes/mediatopic, http://show.newscodes.org/index.html?newscodes=medtop

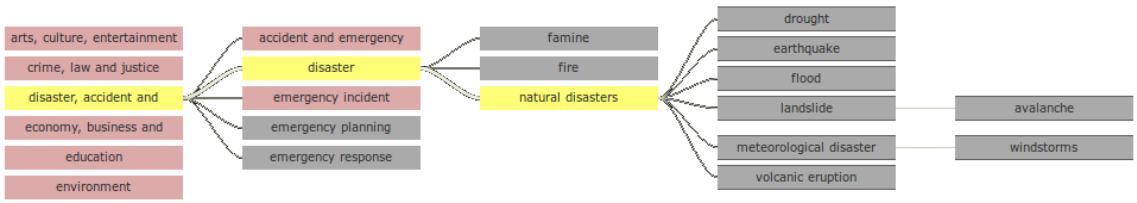


Figure 3: Excerpt of the IPTC Media Topics taxonomy.

Date. The article has to be written at most the day following the *point_in_time*, or between *start_time* and *end_time* if the event has a duration.

Location. The location of a Wikidata instance is defined by the properties *country* (P17) and *location* (P276). One of these values must have be mentioned in the AFP article.

Subject. The Wikidata Event Type (WET) and the title of the instance define a list of keywords relevant to the subject of the article. The similarity score is then the sum of a IMT-based *tf.idf* of the keywords occurring also in the AFP article:

$$score(article) = \sum_{t_i \in article} tf.idf_{IMT}(t_i) \times \mathbf{1}_{keywords}(t_i) \quad (1)$$

where

$$\mathbf{1}_{keywords}(t_i) = \begin{cases} 1 & \text{if } t_i \in WET \\ 0 & \text{otherwise} \end{cases}$$

and

$$tf.idf_{IMT}(t_i) = tf_{IMT}(t_i) \times log\frac{N}{df_{IMT}(t_i)} \quad (2)$$

with t_i the i^{th} token in the article, $df_{IMT}(t_i)$ the number of IMTs associated with documents containing t_i, $tf_{IMT}(t_i)$ the number of occurrences of t_i in the articles sharing the same IMT, and N the number of IMTs. If the AFP article has more than one IMT, then the highest tf_{IMT} for each word is considered.

The weights are based on IPTC Media Topics. Hence, they increase the relevance of a token in a particular context. For instance, "Police" will be more relevant in an event regarding a crime compared to an event regarding an earthquake, even if the token is present in both articles.

An event is considered to be about the same subject as a Wikidata instance if this score is higher than a threshold. Empirically, this threshold is set to 0.04 and roughly corresponds to two tokens in common between the article and the list of keywords.

4.1.2 Evaluation. After this processing, 97,160 articles have been mapped to 8,350 Wikidata instances from 933 Wikidata event types (out of 42k Wikidata events in the considered temporal range). Note that it is neither necessary nor desirable that all articles be mapped to Wikidata instances. Most AFP articles do not relate to events that are meant to end up in Wikidata (*i.e.*, a political speech, a reaction to an event or a trivial news event) and a lot of Wikidata events are not described by any press agency (book publication, local festival or cultural event, TED talks…).

Table 1: Evaluation of the mapping between AFP articles and Wikidata events.

# sentences	Precision	Recall	F1-score
3	1.00	0.67	0.80
5	0.99	0.71	0.83
all	0.96	0.75	0.84

In order to evaluate the quality of this mapping, we manually built a set of 406 pairs *(article, Wikidata instance)* based on 88 instances explicitly linked to a WikiNews page[8] in Wikidata. Only a very small part of Wikidata events are linked to a WikiNews page[9] but we used them to ensure that only "media-worthy" Wikidata events are considered in our manual evaluation. We share this gold standard for further comparison[10].

We report in Table 1 the precision (correct mapping rate), recall (1−missed mapping rate) and F1-measure, when considering the first 3, the first 5 or all the sentences in the AFP articles. These scores show the high quality of the mapping, and consequently, only a very tiny fraction of incorrect information will be shown to the user. Note that missed mappings do not prevent the event to be queried and visualized by our search engine. We only miss the link between the structured and the unstructured data.

4.2 Schema Clustering

As depicted in Table 2, Wikidata event types are often fine-grained and the subclass hierarchy can vary in terms of quality and depth. We seek more coarse-grained event categories for the relevance and robustness of our news classification and potential filters for our search engine. Indeed, from a human perspective, several WETs (e.g. NATO summits and G20 summits) share the same or a very similar structure and clustering them together will make the classification process easier, as well as simplify the interaction with the user within the search engine interface.

To do so, we adopted a hierarchical clustering method based on 3 similarity measures. Each similarity is based on a different representation of the Wikidata Event Type.

- **Label representation**: Even if composed of only a few words (see Table 2), the labels of the Wikidata Event Type (WET) can be a good clue for deciding whether two clusters are similar or not. For instance, the labels "*Election in UK*",

[8] https://en.wikinews.org
[9] Note that our automatic mapping could be used to automate this Wikidata/WikiNews mapping.
[10] https://github.com/crudnik/asrael

Table 2: Examples of Wikidata Event Types for three target schemas.

Target schema	Related Wikidata event types
Election	Bundestag election, direct election, Elections in Saudi Arabia, ...
Plane crash	Aviation accident, plane crash, mid-air collision, ...
Summit	NATO summit, G20 summit, ...

Table 3: Generic tokens.

NER label	New mentions
GPE	geopolitical entity
ORG	organization
PERSON	person
NORP	nationality
DATE	date

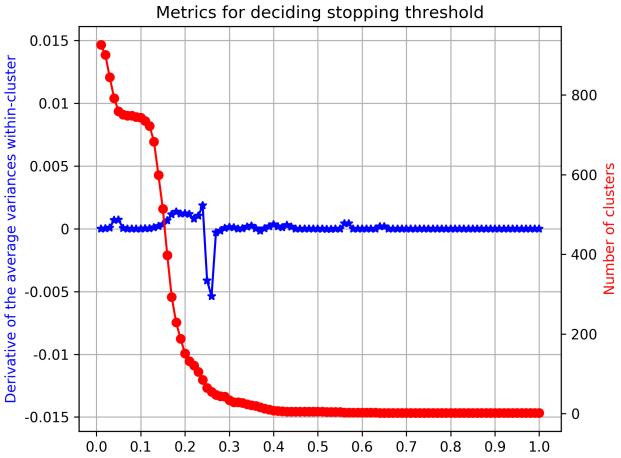


Figure 4: Number of clusters and derivative of the distance within clusters as a function of the similarity threshold to cut the dendrogram.

"*Election in France*" and "*Election*" should be clustered together. However, this approach needs to exclude mentions of organizations or locations, as in "*earthquake in New Zealand*" compared to "*New Zealand general election*", which are arguably not similar in terms of event type. Therefore, we replace mentions of named entities with generic tokens using the spaCy[11] named entity recognition system following the Table 3 (e.g. "*France*" is replaced by "GEOPOLITICAL ENTITY", "*jan-7*" by "DATE").

The representation of the labels $R_l(T)$ of a WET T is the mean of the word2vec [8] vectors of their words.

$$R_l(T) = \underset{w \in label(T)}{mean} (w2v(w)) \quad (3)$$

where $label(T)$ is the set of words in T's label and $w2v(w)$ is the word2vec representation of the word w.

- **Content representation**: this representation is based on the content of the articles. A *WET document* is built by concatenating all the articles mapped to this WET at the previous step (Section 4.1). The content representation $R_c(T)$ is then a vector of all words t_i weighted by their $tf.idf_{WET}$, computed as follows:

$$tf.idf_{WET}(t_i) = tf_{WET}(t_i) \times log \frac{M}{df_{WET}(t_i)} \quad (4)$$

where $tf_{WET}(t_i)$ is the number of occurrences of the term in the WET document, $df_{WET}(t_i)$ is the number of WET documents containing the term, and M is the total number of $WETs$ in our dataset.

- **IPTC Media Topic representation**: in order to improve and to facilitate the stopping decision of the clustering, we add a feature based on the IMTs (see Section 3.3).
As each AFP article is associated with one or several IMTs, the mapping described in Section 4.1 provides also a mapping between a WET and a list of IMTs. We interpret these codes as a new vocabulary describing the WET and we use again a *tf.idf* representation of this new vocabulary. The representation $R_{imt}(T)$ is a sparse vector of size equal to the total number of IMTs present in the corpus, where, for each IMT imt:

$$tf.idf_T(imt) = tf_T(imt) \times log \frac{M}{df_T(imt)} \quad (5)$$

where $tf_T(imt)$ is the number of articles labeled by the IMT imt that have been mapped to a Wikidata event of type T, $df_T(imt)$ is the number of WETs mapped with at least one article with label imt, M is the total number of $WETs$ in our dataset.

We use these three representations to compute the following similarity between two WETs T_i and T_j:

$$sim(T_i, T_j) = \alpha \times \cos(R_l(T_i), R_l(T_j)) \quad (6)$$
$$+ \beta \times \cos(R_c(T_i), R_c(T_j)) \quad (7)$$
$$+ \gamma \times \cos(R_{imt}(T_i), R_{imt}(T_j)) \quad (8)$$

where cos is the cosine similarity measure and the weights α, β and γ are empirically set as 0.38, 0.57, 0.05 respectively.

Note that each of the "label", "content" and "IMT" representations has a different role. Increasing α gives a higher weight to very short texts, which are generally difficult to compare [4]. This would, for example, make closer labels such as 'strike', 'general strike' and 'military strike', which would decrease the quality of the clustering. Content representation is based on longer pieces of text, but increasing β would give a higher weight to potential errors in the mapping. Finally, the IMT similarity is quite categorical compared to the other ones. We want to use it only as an help for choosing when to stop the clustering, which explains the low value of γ in the global similarity score. The importance of this balance between all sources of information is at the same time a strength and a limitation of the method.

Our agglomerative hierarchical clustering procedure is based on the Ward's method. To cut the resulting dendrogram we used a

[11] http://spacy.io/

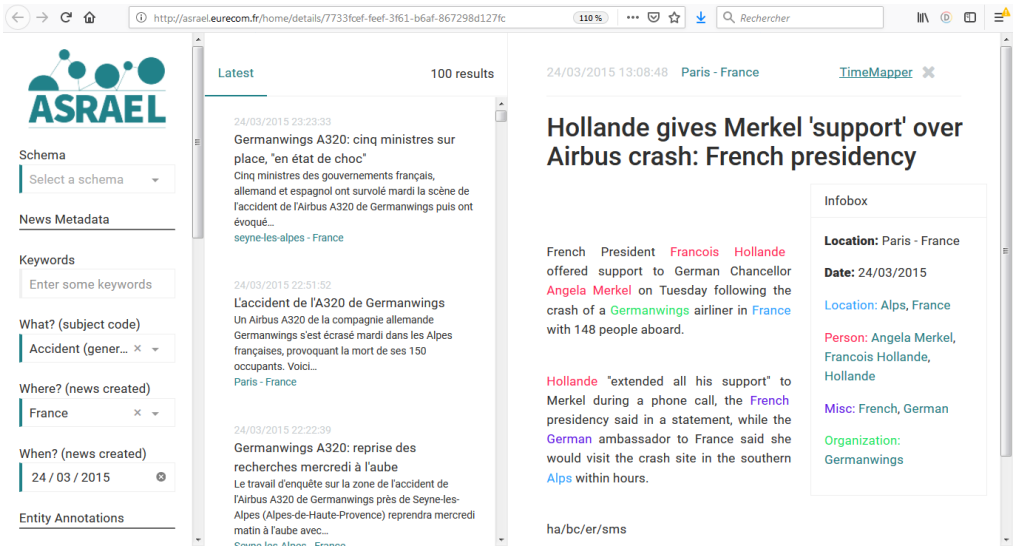


Figure 5: Search engine filtering articles describing plane crash events occurring in France on 24 March 2015.

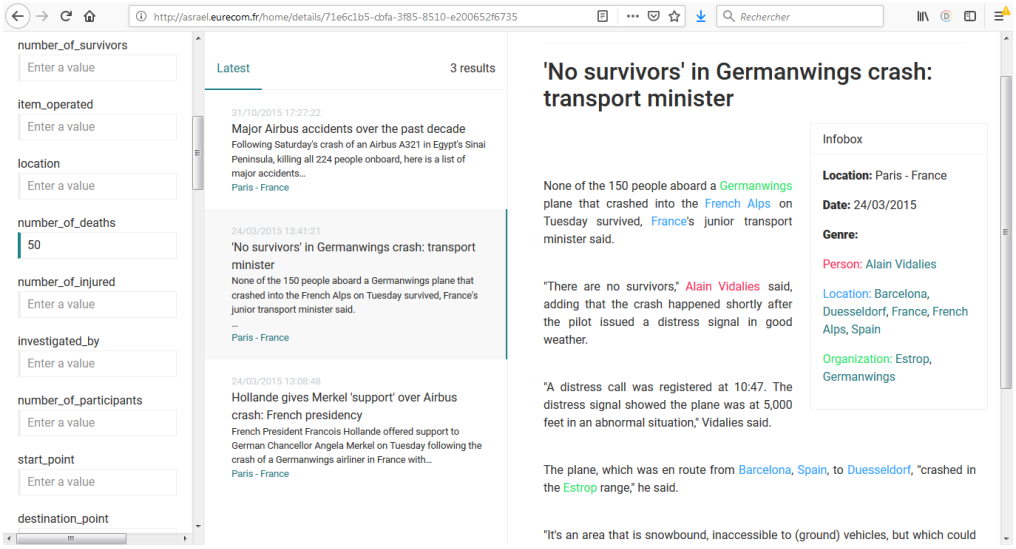


Figure 6: Search engine filtering articles describing plane crash events having caused at least 50 victims.

threshold defined by the elbow method, aiming at finding at which number of clusters the marginal gain of variance will start dropping. According to the graph of Figure 4, the best threshold is 0.23. This leads from 933 initial Wikidata event types to 119 clusters in total. Some extracts of the dendrogram are available in Figure 7.

We observe that this clustering step enables to group together natural disasters, or summits (NATO, G8, ...) into coherent clusters. We also obtained several Election clusters, with the three main ones which seem related to legislative elections, parliamentary elections and general elections.

We empirically evaluated the quality of the clusters for choosing the parameters of our model. Building a protocol for a formal evaluation of this step is a future work. As for most clustering tasks, there is no unique good solution and an automatic, reproducible evaluation seems difficult to set up.

4.3 Automatic Semantic Annotation of News Articles

Our objective is to annotate semantically AFP news articles leveraging Wikidata structured data describing events being told in those articles. We distinguish Wikidata properties that have textual values from the ones having numerical values.

4.3.1 Entity Annotation. When Wikidata properties have a textual value, our goal is to search whether this value is present or not

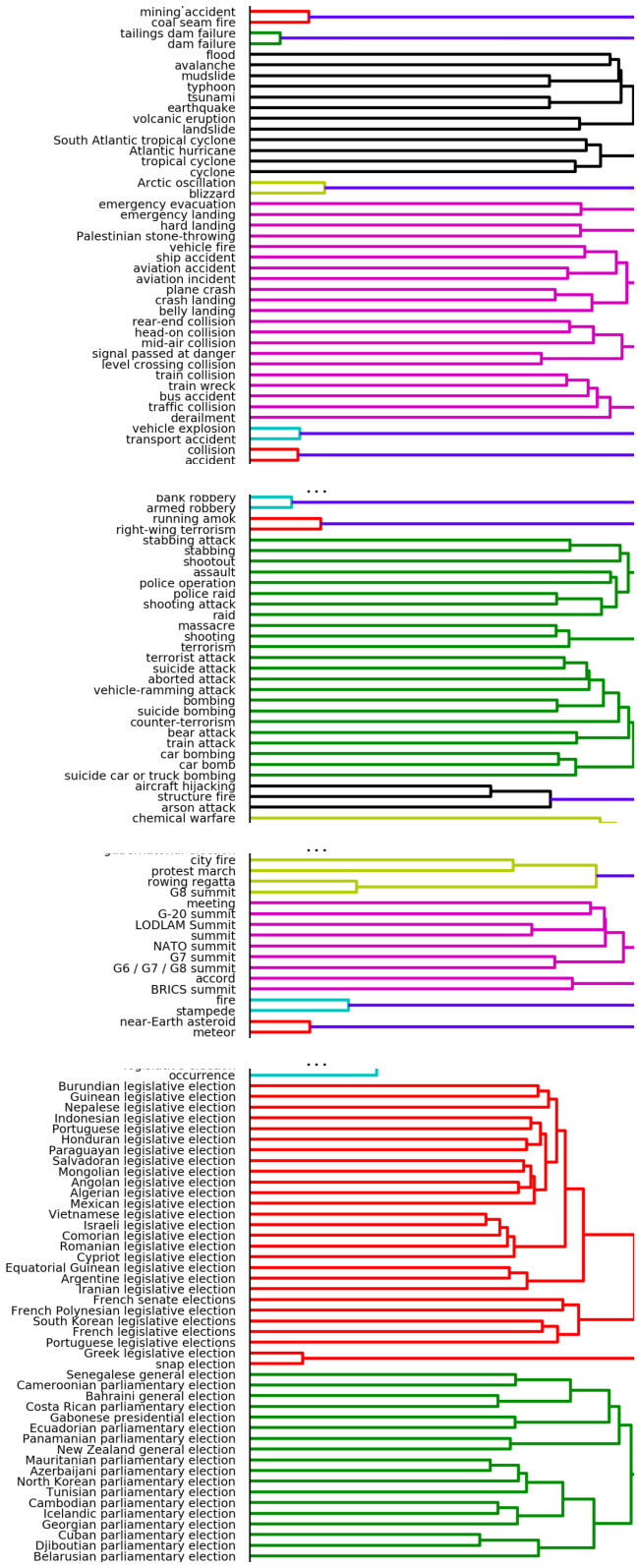


Figure 7: Extracts of the dendrogram issued after clustering.

in the news article. We also use DBpedia redirections to collect different variations of the mention to search in the text.

4.3.2 Quantity Annotation. When Wikidata properties have a numerical quantity, the problem of finding this information in the news article is much harder. The news articles come as a continuous stream and some reported information generally evolves over time, such as the death toll after a deadly accident. Consequently, a proper semantic annotation of a news article should not always consider the exact quantity value indicated in Wikidata. Therefore, for quantitative values, we introduce some flexibility and a quantity in the text (float or integer) is added to a candidate list of annotations if it is in the +/- 10% range of a property and if it is in the first five sentences of the article (supposedly, the earlier in the text, the most relevant to the article main subject). These candidates are then ranked according to the relevance of the semantic context of the numerical value in the text (considering that the context expresses the type of the attribute). Consequently, given two quantities, the more the context (e.g. "9 were killed on Saturday") is similar with the property type (e.g. *number_of_killed* rather than *magnitude_on_Richter_scale*), the more probable the quantity is to be linked with this property. The news article is then annotated with the wikidata property and the most relevant quantity.

4.3.3 Serializing the Annotations. We represent both the news articles metadata and the semantic annotations in RDF. We first convert the AFP news article metadata encoded in NewsML in RDF using the rNews vocabulary. For example, the metadata associated with the news article described by Listing 1 indicates that this story was created on 24/03/2015 with the English headline *'No survivors' in Germanwings crash: transport minister.*

```
<http://asrael.eurecom.fr/news/71e6c1b5-cbfa-3f85-8510-e200652f6735>
  a  rnews:Article ;
  rnews:dateCreated "2015-03-24T12:41:21Z"^^xsd:dateTime;
  rnews:headline  "'No survivors' in Germanwings crash: transport minister"@en ;
  dc:subject  <http://cv.iptc.org/newscodes/subjectcode/03013000>,
    <http://cv.iptc.org/newscodes/subjectcode/04015000>,
    <http://cv.iptc.org/newscodes/subjectcode/03010000>,
    <http://cv.iptc.org/newscodes/subjectcode/04000000>,
    <http://cv.iptc.org/newscodes/subjectcode/03010003>,
    <http://cv.iptc.org/newscodes/subjectcode/04015001>,
    <http://cv.iptc.org/newscodes/subjectcode/03000000>;
  schema:keywords   "minister", "aviation", "accident",
     "Germany", "Spain", "survivors", "France" .
```

Listing 1: Semantic annotation of a news article

The event being described in this news article exists in Wikidata as `Q19671417`. We create an instance of the `schema:Event`[12] which is about this news article. In this article, the number of dead people (150) is correctly found (Listing 2). The schema *S34* is one of the schema output of the clustering phase described in the section 4.2.

```
<http://asrael.eurecom.fr/news/71e6c1b5-cbfa-3f85-8510-e200652f6735>
   rnews:about  <http://asrael.eurecom.fr/event/71e6c1b5-cbfa-3f85-8510-
         e200652f6735> .

<http://asrael.eurecom.fr/event/71e6c1b5-cbfa-3f85-8510-e200652f6735>
   a    schema:Event , wd:Q750215 , rnews:Concept ;
   rdfs:label  "'No survivors' in Germanwings crash: transport minister" ;
   dc:identifier "urn:newsml:afp.com:20150324T124135Z:TX-PAR-ENS90:5" ;
   owl:sameAs wd:Q19671417 ;
   wdt:P1120  "150" ;
   wdt:schema "S34" .
```

Listing 2: Semantic annotation of a news article

[12]The `schema` prefix refers to the Schema.org vocabulary.

4.3.4 Annotation Dataset. As a result, we created annotations associated with 370 properties extracted from Wikidata. This dataset can easily be used for a relation extraction task, with a distant supervision system. The evaluation of this automatic annotation is part of our future work. Note that this step is not necessary to build the search engine described hereafter.

4.4 Search Engine

We load all RDF annotations in a triple store using the Openlink Virtuoso software. The full text of the news article is also indexed in the triple store. We then developed a user interface that performs SPARQL queries to provide views on the data. The Figure 5 depicts the view of the news article[13]. On the top right, we show an infobox composed of the main named entities extracted in the article using the ADEL system [11]. The Figure 6 depicts the view of the news article[14]. On the left panel, the user has selected the schema *S34* corresponding to crash accident. Therefore, a set of additional properties are automatically suggested as new filters, such as the number of victims. The user has entered the value 50 and the search engines retrieves the news articles that describe crash accidents that have yielded at least this number of victims.

5 CONCLUSION AND FUTURE WORK

In this paper, we develop an event-based search engine capable to query both the structure data of knowledge bases and the unstructured textual content of news articles. This facilitates the navigation through events of the same type and aggregate complementary information about the same event. Furthermore, we produced a semi-automatically annotated text dataset. This dataset could be used as a distant supervision for training an annotation system. This system could then be able to extract the structure of the events from the news article, even if they are not in Wikidata. We also plan to work on a multilingual support for this system.

ACKNOWLEDGEMENTS

This work has been partially supported by the French National Research Agency (ANR) within the ASRAEL Project, under grant number ANR-15-CE23-0018, and the ContentCheck Project, under grant number ANR-15-CE23-0025-01.

[13] http://asrael.eurecom.fr/home/details/7733fcef-feef-3f61-b6af-867298d127fc
[14] http://asrael.eurecom.fr/home/details/71e6c1b5-cbfa-3f85-8510-e200652f6735

REFERENCES

[1] Razvan Bunescu and Marius Paşca. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
[2] Simon Gottschalk and Elena Demidova. 2018. EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In *Extended Semantic Web Conference (ESWC)*.
[3] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence, special issue on Wikipedia and Semi-Structured Resources* (2013).
[4] Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, 515–520.
[5] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2013. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* (2013).
[6] Lijun Lyu and Besnik Fetahu. 2018. Real-time Event-based News Suggestion for Wikipedia Pages from News Streams. In *Wiki Workshop*. Lyon, France, 1793–1799.
[7] Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *16th ACM Conference on Conference on Information and Knowledge Management (CIKM)*. Lisbon, Portugal, 233–242.
[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NIPS)*. 3111–3119.
[9] Arunav Mishra. 2014. Linking Today's Wikipedia and News from the Past. In *7th ACM Workshop on Ph.D Students on Information and Knowledge Management (PIKM)*. Shanghai, China, 1–8.
[10] Arunav Mishra and Klaus Berberich. 2016. Leveraging Semantic Annotations to Link Wikipedia and News Archives. In *38th European Conference on Information Retrieval (ECIR)*. Padua, Italy, 30–42.
[11] Julien Plu, Giuseppe Rizzo, and Raphael Troncy. 2019. ADEL: ADaptable Entity Linking. *Semantic Web Journal* (2019).
[12] Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher Manning, and Daniel Jurafsky. 2014. Event Extraction Using Distant Supervision. In *9th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.
[13] Mitja Trampuš and Blaz Novak. 2012. The Internals Of An Aggregated Web News Feed. In *Conference on Data Mining and Data Warehouses (SiKDD)*.
[14] Giang Binh Tran and Mohammad Alrifai. 2014. Indexing and Analyzing Wikipedia's Current Events Portal, the Daily News Summaries by the Crowd. In *23rd International Conference on World Wide Web (WWW), Demo Track*. Seoul, Korea, 511–516.
[15] Raphaël Troncy. 2008. Bringing the IPTC News Architecture into the Semantic Web. In *7th International Semantic Web Conference (ISWC)*. Karlsruhe, Germany, 483–498.
[16] Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics* 9, 2 (2011), 128–136.
[17] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

Finding Prerequisite Relations using the Wikipedia Clickstream

Mohsen Sayyadiharikandeh
Indiana University, Bloomington
Bloomington, Indiana, USA
msayyadi@indiana.edu

Jonathan Gordon
Vassar College
Poughkeepsie, New York, USA
jgordon@vassar.edu

José Luis Ambite
USC Information Sciences Institute
Marina del Rey, California, USA
ambite@isi.edu

Kristina Lerman
USC Information Sciences Institute
Marina del Rey, California, USA
lerman@isi.edu

ABSTRACT

The increased availability of online learning resources in the form of courses, videos, and tutorials has created new opportunities for independent learners, but it has also increased the difficulty of planning a course of study. Where should the learner start? What should the learner know before tackling a new course? Manually identifying these prerequisite relations between learning resources or concepts is expensive in terms of time and expertise, and it is particularly difficult to do so for new or rapidly changing areas of knowledge. To address this challenge, we present a new method for identifying prerequisite relations based on naturally occurring data, namely the navigation patterns of users on the Wikipedia online encyclopedia. Our supervised learning approach shows that the navigation network structure can be used to identify dependencies among concepts in several domains.

KEYWORDS

Prerequisite; Clickstream; Wikipedia; Concept graph

ACM Reference Format:
Mohsen Sayyadiharikandeh, Jonathan Gordon, José Luis Ambite, and Kristina Lerman. 2019. Finding Prerequisite Relations using the Wikipedia Clickstream. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316753

1 INTRODUCTION

Self-directed learners can benefit greatly from the scientific and technical training resources available online. However, it can be challenging to organize these resources into a suitable educational plan. In traditional education, whether in the classroom or in textbooks, concepts are taught in a sequence determined by an expert's understanding of the domain. Self-directed online learners, however, may not know where to begin. For example, if your goal is to understand recursive neural networks, you may not know that you first need to understand several more basic concepts in mathematics and machine learning.

In this task, learners can be guided by a *concept graph*, which links particular educational resources – such as online courses, videos, and tutorials – to the concepts they cover and links each concept to any more basic, prerequisite concepts. We consider a concept A to be a prerequisite of concept B if A is necessary or significantly helpful for understanding B. (Other relations, such as similarity or hyponymy, can hold between concepts without either being a prerequisite of the other.) A concept graph can be used directly by a learner to explore the conceptual structure of the domain, like looking at a map, or it can be used by applications to recommend particular learning resources based on the prerequisites between concepts, as in the generation of structured reading lists [5].

Manually constructing a concept graph is time-consuming and requires significant domain knowledge, motivating automatic methods. We introduce an approach that infers prerequisite relations between concepts based on the navigation of users on Wikipedia. We treat encyclopedia articles as identifying concepts – with varying levels of granularity – and train a classifier to predict whether concept A is a prerequisite of concept B based on features computed from a graph of "clickstream" data, where articles are connected by weighted edges that indicate the number of times users followed links from one article to the other. Intuitively, users visiting an article are interested in learning about that concept, and they will follow links to other articles they believe will support that objective. Thus, navigation tends to flow from a concept to its prerequisites.

For training and testing, we use two sets of gold-standard prerequisite data. One is an existing set, which covers several domains and uses crowdsourced judgments of whether a prerequisite relation holds between a pair of Wikipedia articles [15]. The other is derived from a large, expert-generated concept graph for machine learning [7], which we semi-automatically map to the most relevant Wikipedia articles.

The contributions of this paper are: (1) a novel approach to prerequisite identification based on the observed behavior of online learners and (2) a new evaluation set for prerequisites based on a semi-automatic mapping of an expert-generated concept graph to Wikipedia.

2 RELATED WORK

Manually curated graphs of prerequisite relations are used to guide learners[1] and plan curricula – i.e., to order learning resources based

[1] For a recent example, see Figure 1.6 in Goodfellow et al.'s 2016 textbook on deep learning [4].

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316753

on the concepts they cover. For instance, Stanford's course CS 234 *Reinforcement Learning* has as a prerequisite CS 229 *Machine Learning*, which lists linear algebra, probability, and statistics as prerequisites.[2] Several researchers have recently presented approaches to automatically infer these relations between concepts and pedagogical resources. Before presenting our approach, we briefly review these efforts.

Talukdar and Cohen [15] crowdsourced judgments of prerequisite relations, and then employed a maximum-entropy classifier with three types of features: (1) derived from Wikipedia's link graph, (2) related to Wikipedia users' edits, and (3) related to article content. Their classifier achieved 58.82% accuracy. Liang et al. [10] proposed a link-based metric for measuring prerequisite relations among Wikipedia concepts. They compute prerequisites based on reference distance (RefD), where Wikipedia links serve as "reference relations" among concepts. (Reference distance is asymmetric, so A and B cannot both be prerequisites of each other.) They evaluate on Talukdar and Cohen's prerequisite data and on their own gold-standard dataset made by crawling university websites and mapping to Wikipedia concepts. Since this new evaluation set finds prerequisites at the level of entire courses, they are significantly more coarse-grained than the Metacademy-based data set we introduce (e.g., a concept would be *Machine learning*, not *Markov chain Monte Carlo*).

Liang et al. [10] re-implemented Talukdar and Cohen's maximum-entropy method, for which they report a higher average accuracy of 60.4%, which may be due to the use of newer Wikipedia data. They also introduce a simpler method which achieves a higher average accuracy of 61.2%. In later work, Liang et al. [11] studied the applicability of an active learning approach applied to their previous method with some novel features. They employed different query strategies for pool-based active learning and concluded that query-by-committee constantly outperforms other methods. They also achieved higher accuracy compared to other related work.

Wang et al. [16] use the order of topics in textbook tables of contents to extract concept maps by jointly optimizing the extraction of key concepts with corresponding Wikipedia articles and the identification of prerequisite relations. They evaluate their method by building concept maps from six textbooks and having domain experts evaluate the results. A limitation of this approach is that it requires concepts to have already been ordered by an expert in a textbook. For many domains, appropriate textbooks may not be readily available, or a concept graph might require combining the orderings from multiple textbooks.

Medio et al. [13] considered predicting prerequisite relations between "learning objects" on Coursera. For each learning object, they find a set of related Wikipedia articles and, using Coursera's gold-standard prerequisites between courses, they train a classifier using textual and hyperlink features of the matched Wikipedia articles. Given that many learning resources can describe the same concepts, we are interested in inferring prerequisite relations among concepts and using these to form learning plans rather than predicting prerequisite relations among particular resources.

[2] http://exploredegrees.stanford.edu/schoolofengineering/computerscience

Gordon et al. [6] present a novel model where each concept – a latent Dirichlet allocation (LDA) topic learned from a corpus of scientific articles – is linked by prerequisite relations as a step in building a concept graph to support the automatic generation of personalized reading lists. Using this formulation of concepts as probability distributions over words, they introduced an information-theoretic view of prerequisite relations based on cross entropy and information flow. An advantage of this work is that it does not rely on an external reference source like Wikipedia, but a limitation is that the results of unsupervised topic discovery are harder for humans to interpret than a Wikipedia article.

2.1 Our Approach

In this paper, we use Wikipedia articles as concepts, which can be linked to learning resources, such as online courses, videos, or documents, e.g., by using Explicit Semantic Analysis [3]. While there are limitations to the use of Wikipedia articles as concepts, such as the problem of identifying sub-articles describing a facet of a more general concept [12], each article has a clear interpretation, and the set of articles has broad coverage.

Rather than use the textual or hyperlink network features of previous work, we use a new data source: the Wikipedia Clickstream, consisting of actual navigation of learners among articles. While we do not assert that this navigation replaces the information that can be learned from other features, we investigate how much it can contribute toward the problem of prerequisite identification.

Like previous work in this area, we use the Talukdar and Cohen [15] Wikipedia evaluation set. However, like Liang et al. [10], we see the need for an additional, naturally occurring evaluation set – albeit with more specific concepts than they used – leading us to create mappings from Metacademy to Wikipedia. These data sets are described in more detail in the next section.

3 DATA SETS

We use three sources of data: the Wikipedia Clickstream and two sets of prerequisite relations used for training and testing. The first is Talukdar and Cohen's multi-domain crowdsourced judgments, used by most previous work, and the second is a new data set based on Metacademy's expert-enumerated prerequisite relations, which were created to guide learners rather than to evaluate research.

3.1 Wikipedia Clickstream

The Wikipedia Clickstream [17] consists of data sets containing counts of (referrer, resource) pairs extracted from the user request logs of the English, Farsi, and Arabic editions of Wikipedia. The data is divided into months of clicks, starting with January 2015, and includes only pairs of articles with more than 10 clicks. The clickstream data can be seen as giving a weighted network of articles [9], where the weights are human navigation through a popular network of learning resources. For research projects, it is a large network; it includes 1.3 million nodes and 22 million edges for January 2017 alone. In addition to navigation between articles, it includes inbound clicks from web sites such as google.com.

We compute our features using the English Wikipedia clickstream data sets that were released for January and February 2015, six months in 2016, and January 2017. We filtered these data sets to

only those links where both articles are in the set of concepts under consideration (articles predicted to be relevant by the automatic mapping of Metacademy concepts or those articles chosen for inclusion in the CMU prerequisite data set, described in the following sections). Note that pairs retrieved at this step include not only positive examples of prerequisite pairs but also (probable) negative examples, where both concepts are of interest in the domain but are not known to have a prerequisite relation.

3.2 CMU Prerequisite Data

The CMU prerequisite data [15] covers five domains: Global Warming, Meiosis, Newton's Laws of Motion, Parallel Postulate, and Public-key Cryptography. For each domain, Talukdar and Cohen selected pairs of articles based on their random walk with restart (RWR) scores. At the time the data was collected, every pair of articles (d, d') in their data set had a hyperlink from d to d'. These pairs were then presented to people on Amazon Mechanical Turk, who were asked whether (1) d' is a prerequisite of d, (2) d is a prerequisite of d', (3) the articles are unrelated, (4) the articles are related but there is no prerequisite relation, or (5) don't know. For the Newton's Laws and Global Warming domains, they collected five votes for each of 400 pairs per domain. For the Parallel Postulate and Public-key Cryptography domains, they collected three votes for each of 200 pairs per domain. For Meiosis, they collected three votes for each 400 pairs. A limitation of these data sets is their modest scale, including fewer than 100 positive pairs of prerequisites per domain.

While naturally occurring sources of prerequisites – like Metacademy, described next – consist of only positive examples, the CMU data set includes three types of negative samples: unrelated pairs, pairs that are related but not prerequisites, and when a prerequisite was identified as being in the other direction. Following the processing described by Talukdar and Cohen, we excluded "Don't know" responses, aggregated the votes, and assigned final labels based on majority vote for each pair, breaking ties arbitrarily. We considered the three mentioned types of relations between pairs as negative examples for the final data sets.

3.3 Metacademy Prerequisite Data

Metacademy (metacademy.org) is a free, open-source platform for learning. It covers 487 concepts related to machine learning, which are connected in 1,208 prerequisite pairs (7,947 under transitive closure). Figure 1 shows a sample Metacademy learning plan, where each edge represents a prerequisite relation. To use the expert-generated Metacademy prerequisite relations for the problem of inferring prerequisites among Wikipedia concepts, we introduce a method to map Metacademy concepts to the closest matches in Wikipedia's semantic space.

Finding Possible Wikipedia Matches. Our initial set of possible matches comes from querying a search engine, since this lets us exploit significant relevance engineering, including features based on human intelligence (since search engines generally take human clicks into account when ranking results). In particular, we searched DuckDuckGo for "*Metacademy concept* + wiki page" and excluded results with URLs that were not from Wikipedia.[3] The granularity of concepts on Metacademy sometimes differs from that of articles on Wikipedia, so some concepts in learning plans do not have exactly corresponding Wikipedia articles, either because they are too specific (e.g., *Higher-order partial derivatives* or *Linear regression closed-form solution*) or because they are too general (e.g., *Expectation and variance*).

Predicting Relevance of Matches. For a sample of 100 Metacademy concepts, 763 search results were annotated as 0 (not related), 0.5 (related), or 1 (exact match, either for the whole article or a section of it). The annotation was performed by two of the authors, with a Pearson correlation of 0.75 ($p < 0.001$). We found that about 89% of samples that were labeled as 1 by both annotators also have rank 1 in search results. So, a simple baseline classifier predicts that the first result for every query is a correct match.[4] To distinguish which of the additional results are likely to be related (0.5) vs not (0), we use the "Category" of the Wikipedia articles: An article whose set of categories overlaps with the categories of the rank-1 result for that query will be labeled as 0.5. If there is no overlap, then the article is likely to be a spurious match. Based on manual inspection, we included two levels of ancestor categories when determining overlap.

As an example, if we search for "d separation", three Wikipedia articles are found in the first page of results: *Bayesian network*, *M-separation*, and *Separation of powers*. Our automatic method labels *Bayesian network* as 1, since it is the top-ranked search result. It labels *M-separation* as 0.5, since it has the common category *Graphical models*, and it labels *Separation of powers* as 0.

While this is a simple method for predicting the relevance of Wikipedia results, it has a 0.7 correlation ($p < 0.001$) with the average of the human annotators – nearly as high as the inter-annotator agreement. We sampled an additional 30 Metacademy concepts for annotation by one of the authors as a validation set. The predictions had a 0.72 correlation with the human annotations for the validation set. Therefore, we used the results of this simple method for finding related Wikipedia articles.

We searched for 487 Metacademy concepts on DuckDuckGo and used our semi-automatic method for the rest of the unannotated matches (results of 487, minus 100 queries). Focusing on matches labeled 0.5 or 1, we retrieved 776 distinct Wikipedia articles (1,091 total matches). For 1,208 learning pairs we get 3,419 pairs of Wikipedia concepts.

Our semi-automatic mapping of Metacademy concepts to Wikipedia is being released at https://doi.org/10.6084/m9.figshare.7799774 and can be used to evaluate other work on prerequisite discovery. Our method for mapping concepts to Wikipedia's semantic space is general, and we expect it can be applied to gold-standard concept graphs that may exist for other domains.

[3] Three additional kinds of searches were also tried, including the basic query, searching with the "site:" restrictor, and searching on Wikipedia itself, but these had lower precision.

[4] This was the top-splitting rule learned by a decision tree classifier trained on the results of the four methods of search we tested. However, the decision tree was less good at predicting 0.5 and 0 annotations, leading us to manually produce the rules described.

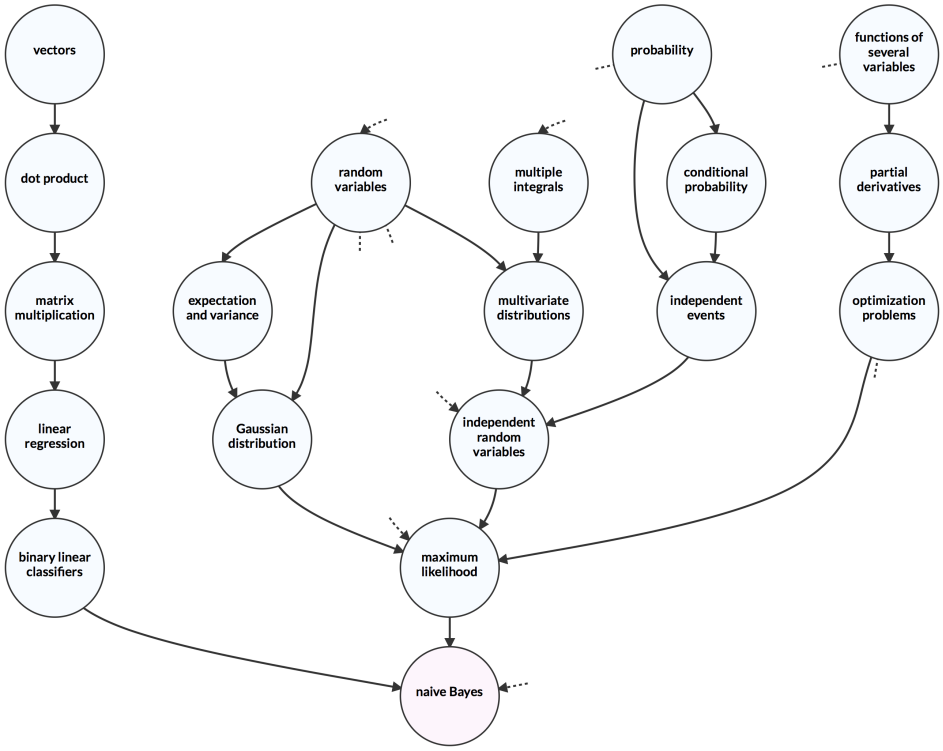


Figure 1: Sample learning plan from Metacademy

3.4 Final Data Sets

The final data sets used for the prerequisite prediction experiments described in the next section consist of the entries from the aggregated Wikipedia clickstream data, where both articles belong to the set of concepts from the CMU or Metacademy prerequisites.

For the Metacademy data, we also compute sets of gold-standard prerequisites based on transitive closure. The transitive closure deals with the fact that, much as concepts can be enumerated at different levels of granularity, prerequisite relations may include intermediate dependent concepts or ignore them, e.g., we can say that *Hidden Markov model* depends on *Stochastic process* or we can say that *Hidden Markov Model* depends on *Markov chain*, which depends on *Stochastic process*. To compute the transitive closure, if (C_1, C_2) and (C_2, C_3) are prerequisite pairs, then we consider (C_1, C_3) as prerequisites as well.

Since we are dealing with a binary classification problem, the supervised learner needs some number of examples for both classes (prerequisite and not). In prerequisite relation identification, it is typically easy to find negative examples, while positive examples are scarce. Using the transitive closure can help for domains where we do not have enough labeled positive examples, ameliorating the problem of class imbalance.

While the CMU data set contains both positive and negative labeled pairs of concepts, Metacademy only contains positive examples, and we require some negative samples for our final data set. As a simple technique for generating negative samples, we make a closed-world assumption: Any example of navigation in the clickstream involving a concept from the domain that is not known to have a prerequisite relationship is taken as a negative example.

Table 1: Class distributions of Metacademy prerequisite data

			Transitive closure	
Data Set	Original MA	Balanced MA-bal	Original MA-TR	Balanced MA-bal-TR
Prerequisites	10%	50%	18%	50%
Non-prereq.	90%	50%	82%	50%

The original Metacademy data set is referred to as MA and the transitive closure, consisting of 14,633 pairs of concepts, as MA-TR. We use sub-sampling of majority class to create MA-bal and MA-bal-TR, which are balanced data sets – ones where there are the same number of observations for each class. The class distributions for the final Metacademy data sets are given in Table 1.

The graph of positive examples from MA is visualized in Figure 2, using the Yifan Hu proportional layout algorithm [8] in Gephi [1]. The large connected subgraph includes most of the relevant Wikipedia pages. Manual inspection of nodes in the graph shows nodes in the small subgraphs are mostly irrelevant results from the automatic matching using a search engine.

4 METHOD

To predict prerequisites, we train supervised classifiers on a set of features defined using the Wikipedia clickstream data for the

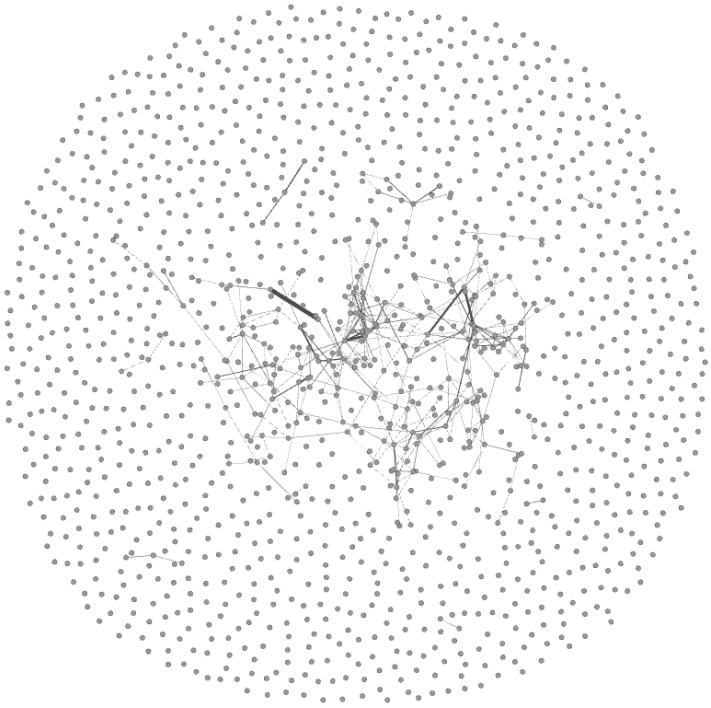


Figure 2: Visualizing positive prerequisite edges associated with Metacademy learning pairs (data set MA)

concepts of interest. In this section, we define the features and discuss our choice of classifiers.

4.1 Features

We focused on user navigation–based features, excluding those related to the content of the articles. We create a directed, weighted graph from the clickstream matches, where the weight of an edge from article A to article B is the total number of clicks. We started with a few predictive features defined from this graph and incrementally added more until we achieved satisfying results. The final features are:

- *Weight.* Total number of clicks from article A to B.
- *Backward weight.* Total number of clicks from B to A.
- *Sum.* Sum of *weight* and *backward weight.*
- *Diff.* Absolute difference between *weight* and *backward weight* features.
- *Sum of all transitions.* Sum of the weights of all outgoing edges from A.
- *Mean of weights.* Average of the weights of all outgoing edges from A.
- *Normalized weight.* Weight divided by sum of transitions.
- *Normalized backward weight.* The backward weight divided by the sum of all transitions for B.
- *Weight greater than mean.* A binary feature indicating whether the *weight* is greater than the *mean of weights.*
- *Backward weight greater than mean.* A binary feature indicating whether the *backward weight* is greater than B's *mean of weights.*

The intuition behind our use of backward navigation features (from an article to its prerequisites) is the idea that users who do not understand concept B will then study concept A if it will help them.

There is a low likelihood that we will have clickstream data (a non-trivial amount of direct navigation in the recorded months) for all pairs of articles in the gold-standard data sets. To improve the coverage of concepts, we also computed the maximum potential navigation (PN) on paths with one or two intermediate nodes. For one intermediate node, this is computed as:

$$\mathrm{PN}_1(A, B) = \max_C[\min(\mathrm{weight}[A \to C], \mathrm{weight}[C \to B])]$$

We did not consider any temporal properties of the available data, e.g., which articles existed or were connected by links for which months. Rather, we summed the clickstream counts for each pair across all available months.

4.2 Classifiers

For each data set, we trained binary classifiers using all of the features, using the scikit-learn [14] package in Python. In initial experiments, we used logistic regression and decision tree classifiers and then added AdaBoost and Gaussian naïve Bayes to compare the performance. We chose logistic regression as a classic classifier suitable for binary classification. We also used Gaussian naïve Bayes since we are dealing with continuous features and naïve Bayes is a simple algorithm used as an initial classification in many experiments. Decision tree is used since it can handle class imbalance better than other classical classifiers and can also produce human understandable rules. AdaBoost is a popular ensemble-learning

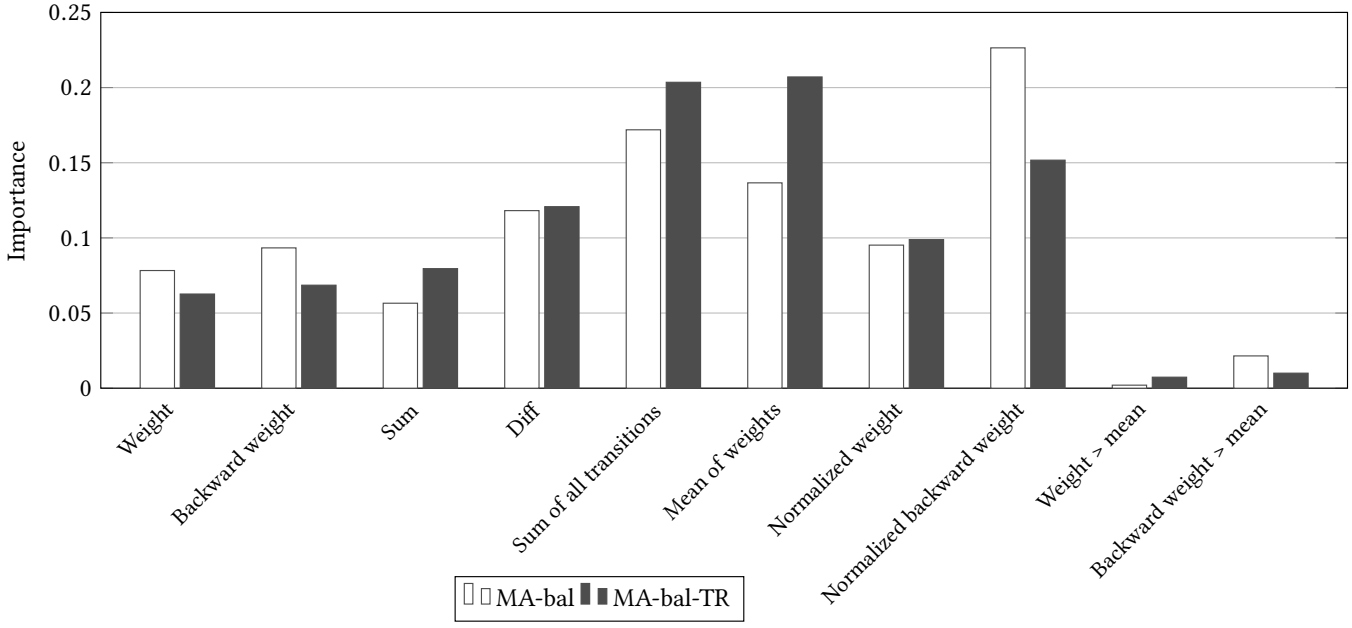


Figure 3: Feature importance for decision tree on MA-bal

method which usually gives more robust models. We set the maximum number of estimators in AdaBoost as 200 and use a decision tree with maximum depth equal to three as base estimators. We used these classifiers with five-fold cross validation and used 80% of the data set as training and 20% for testing.

5 RESULTS

Our goal is to label a pair of concepts as a prerequisite or not a prerequisite. We measure performance of classifiers on the Metacademy and CMU prerequisite data in terms of the accuracy, precision, recall, and F_1 scores. Accuracy shows the average performance of each classifier as the fraction of the pairs that have been successfully labeled.

$$\text{Accuracy} = \frac{\text{Correct predictions (both classes)}}{\text{All samples}}$$

Precision is the fraction of correctly labeled pairs.

$$\text{Precision} = \frac{\text{Prereq. pairs} \cap \text{Retrieved pairs}}{\text{Retrieved pairs}}$$

Recall calculates what fraction of all prerequisites are identified.

$$\text{Recall} = \frac{\text{Prereq. pairs} \cap \text{Retrieved pairs}}{\text{Prereq. pairs}}$$

In Tables 2 and 3, we present the performance of our classifiers for the original and transitive closure sets of prerequisite relations from Metacademy. The AdaBoost and decision tree classifiers, trained over balanced data sets (MA-bal and MA-bal-TR) produce the best results achieving precision, recall, and F_1 scores of around 80%.

Since negative examples constitute 90% of data set MA, a lazy conservative classifier can predict with 90% accuracy simply by predicting every pair of concepts not to be prerequisites. (This is known as the accuracy paradox.) We include this lazy classifier and a random classifier as baselines. In the Metacademy prerequisite data sets MA and MA-TR, the performance of logistic regression and Gaussian naïve Bayes classifiers could not beat the baseline lazy classifier. As expected, logistic regression is highly sensitive to class imbalance, and it ignores the minority class. This is due to the cost function and update rule of logistic regression, where a good model for the majority class can minimize the cost function. The naïve Bayes classifier gives poor performance because we are using a basic version of naïve Bayes in which the class prior biases the predictions toward the majority class.

Performance can be improved by using a classifier that better handles class imbalance, such as decision tree. Decision tree is resilient against class imbalance because it selects the splitting rules based on information gain (or Gini index), which can force both classes to be addressed. The scikit-learn library uses an optimized version of the CART algorithm [2], which tries to induce a tree with largest information gain at each node. Sub-sampling the majority class can also help our classifier to boost the precision and recall in the balanced data sets (MA-bal and MA-bal-TR).

In our experiments, using the transitive closure did not significantly change the performance. This is the case because our gold standard had enough positive pairs and even finding matched clickstream records for a subset of them sufficed to train a good classifier. Especially in the original data set (MA), using transitive closure (MA-TR) could not beat the class imbalance problem in our data set.

Looking at the learned decision tree for MA, the top splitting feature is *Normalized backward weight*. This supports our hypothesis that learners will read an article and, if they do not understand the concept, will navigate to its prerequisite concepts. Figure 3 shows

Table 2: Performance of predictions on the Metacademy datasets MA and MA-bal

	MA				MA-bal			
	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.	F_1
Random Classifier	0.50	0.10	0.50	0.17	0.50	0.50	0.50	0.50
Lazy Conservative Classifier	0.90	0	0	0	0.50	0	0	0
Logistic Regression	0.91	0.03	0.01	0.01	0.55	0.61	0.37	0.46
Gaussian Naïve Bayes	0.89	0.17	0.10	0.12	0.74	0.18	0.29	0.22
Decision Tree	**0.93**	0.61	**0.64**	**0.62**	**0.81**	0.78	0.77	0.77
AdaBoost	0.92	**0.69**	0.13	0.22	0.80	**0.80**	**0.78**	**0.80**

Table 3: Performance of predictions on the Metacademy datasets with transitive closure, MA-TR and MA-bal-TR

	MA-TR				MA-bal-TR			
	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.	F_1
Random Classifier	0.50	0.18	0.50	0.26	0.50	0.50	0.50	0.50
Lazy Conservative Classifier	0.82	0	0	0	0.50	0	0	0
Logistic Regression	0.88	0.12	0.01	0.02	0.56	0.60	0.57	0.59
Gaussian Naïve Bayes	0.86	0.19	0.06	0.09	0.54	0.67	0.16	0.26
Decision Tree	**0.91**	0.62	**0.65**	**0.63**	**0.78**	0.77	**0.79**	**0.78**
AdaBoost	0.90	**0.73**	0.17	0.28	**0.78**	**0.78**	0.78	**0.78**

Table 4: Coverage for the CMU and Metacademy data sets

Coverage with	Newton's Laws	Public-key Cryptography	Global Warming	Parallel Postulate	Meiosis	Metacademy
Direct link	41%	63%	55%	36%	58%	18%
1 intermediate node	55%	72%	78%	61%	82%	30%
2 intermediate nodes	60%	79%	88%	68%	86%	31%

the importance of different features for a decision tree classifier based on Gini index computed on MA-bal and MA-bal-TR.

Table 4 shows the coverage of the Wikipedia clickstream data on different domains of the CMU data set and Metacademy. Including intermediate nodes helps to increase the coverage especially for the CMU data set. The average coverage using direct links is 51%, which increases to 70% while using one intermediate node and 76% while using two intermediate nodes. Note that there are multiple reasons there may not be a match in the clickstream. For the Metacademy data, 15% of the concepts have no matching Wikipedia articles. Another possibility is that there is no direct hyperlink between the Wikipedia articles; using the Wikipedia API, we find that for only 28% of these pairs we do not find in the clickstream there is an associated link in Wikipedia.

Table 5 shows the performance of a decision tree classifier for the CMU data set over the covered concepts. The performance of the classifiers followed the pattern observed for the Metacademy data sets, so we only report the best results from the decision tree, although AdaBoost performed similarly. It is difficult to directly compare our results to previous work using the CMU prerequisite data set. Talukdar and Cohen [15] do not explain the details of their in-domain training approach (since their main focus was out-of-domain), so we cannot mimic the details of their experiment for a fair comparison. However, for the subset of the data set covered by the Wikipedia clickstream, our predications are more accurate than their maximum entropy classifier, as shown in Figure 4. Liang et al. [10] also compare their performance on the CMU data set, using out of domain training. Since our approach uses in-domain training, we cannot directly compare our results to theirs.

Table 5: Performance of our decision tree classifier on the five domains of the CMU data set

	Acc.	Prec.	Recall	F_1 Score
Newton's Laws	0.83	0.81	0.87	0.84
Public-key Cryptography	0.82	0.81	0.85	0.83
Meiosis	0.85	0.84	0.86	0.85
Parallel Postulate	0.81	0.82	0.82	0.82
Global Warming	0.87	0.86	0.88	0.87

6 CONCLUSION

To help learners follow a coherent path through a knowledge domain, we want to infer when one concept is a prerequisite of another. We have described a new approach to this problem that exploits the "clickstream" of human navigation among articles on Wikipedia. In particular, we find that an important feature is the backward

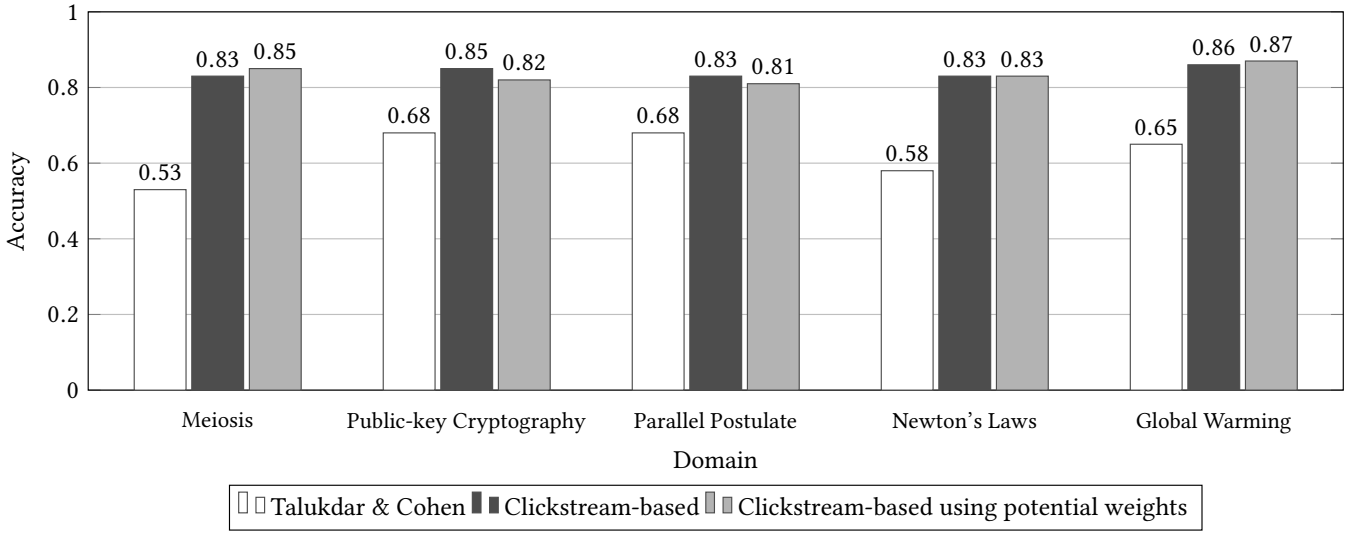


Figure 4: Comparison of our accuracy with Talukdar and Cohen [15] for the subset of the CMU data set covered by the Wikipedia clickstream

navigation from a more advanced concept to one of its prerequisites. We evaluate our classification methods over an existing gold standard of prerequisites crowdsourced by researchers at CMU and over a new data set of expert-generated prerequisites from Metacademy that we map into Wikipedia's semantic space. Training decision tree and AdaBoost classifiers over balanced datasets, we obtain precision, recall and F_1 scores of around 80% on this new Metacademy dataset and F_1 measures from 82% to 87% on the CMU dataset. We hope that these new techniques based on navigation data will enable self-directed learners to take a greater advantage of the vast amounts of learning materials available on the Web.

ACKNOWLEDGMENTS

This work is supported by the National Institutes of Health under Grant U24ES026465. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health.

REFERENCES

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (*ICWSM*). AAAI Press, Palo Alto, CA, USA, 361–2. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[2] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Chapman & Hall / CRC, New York.

[3] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (*IJCAI*). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1606–11.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press, Cambridge, MA, USA. https://www.deeplearningbook.org.

[5] Jonathan Gordon, Stephen Aguilar, Emily Sheng, and Gully A. P. C. Burns. 2017. Structured Generation of Technical Readings Lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA*). Association for Computational Linguistics, Copenhagen, Denmark, 261–70. https://doi.org/10.18653/v1/W17-5029

[6] Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully A. P. C. Burns. 2016. Modeling Concept Dependencies in a Scientific Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*ACL*). Association for Computational Linguistics, Berlin, Germany, 866–75. https://doi.org/10.18653/v1/P16-1082

[7] Roger Grosse and Colorado Reed. 2013. Metacademy. https://metacademy.org.

[8] Yifan F. Hu. 2005. Efficient and high quality force-directed graph drawing. *The Mathematica Journal* 10 (2005), 37–71. Issue 1.

[9] Daniel Lamprecht, Denis Helic, and Markus Strohmaier. 2015. Quo Vadis? On the Effects of Wikipedia's Policies on Navigation. In *Wikipedia, a Social Pedia: Research Challenges and Opportunities: Workshop at the Ninth International Conference on Web and Social Media.* AAAI Press, Palo Alto, CA, USA, 64–6. https://aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10653

[10] Chen Liang, Zhaohui Wu, Wenyi Huang, and C. Lee Giles. 2015. Measuring Prerequisite Relations Among Concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Lisbon, Portugal, 1668–74. https://doi.org/10.18653/v1/D15-1193

[11] Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C. Lee Giles. 2018. Investigating Active Learning for Concept Prerequisite Learning. In *Proceedings of the Eighth Symposium on Educational Advances in Artificial Intelligence* (*EAAI*). AAAI Press, Palo Alto, CA, USA, 7913–9.

[12] Yilun Lin, Bowen Yu, Andrew Hall, and Brent Hecht. 2017. Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing* (*CSCW*). ACM, New York, NY, USA, 2052–67. https://doi.org/10.1145/2998181.2998274

[13] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. 2016. Automatic Extraction of Prerequisites Among Learning Objects Using Wikipedia-Based Content Analysis. In *Proccedings of the 13th International Conference on Intelligent Tutoring Systems* (*ITS*), Alessandro Micarelli, John Stamper, and Kitty Panourgia (Eds.). Springer International Publishing, Switzerland, 375–81. https://doi.org/10.1007/978-3-319-39583-8_44

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–30.

[15] Partha Pratim Talukdar and William W. Cohen. 2012. Crowdsourced Comprehension: Predicting Prerequisite Structure in Wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP.* Association for Computational Linguistics, Montréal, Canada, 307–15.

[16] Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C. Lee Giles. 2016. Using Prerequisites to Extract Concept Maps from Textbooks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (*CIKM*). ACM, New York, NY, USA, 317–26. https://doi.org/10.1145/2983323.2983725

[17] Ellery Wulczyn and Dario Taraborelli. 2017. Wikipedia Clickstream. https://doi.org/10.6084/m9.figshare.1305770.v22

How Partisanship and Perceived Political Bias Affect Wikipedia Entries of News Sources

Khonzodakhon Umarova
Department of Computer Science
Wellesley College
kumarova@wellesley.edu

Eni Mustafaraj
Department of Computer Science
Wellesley College
emustafaraj@wellesley.edu

ABSTRACT

Increased polarization and partisanship have become a consistent state of politics, media, and society, especially in the United States. As many news publishers are perceived as "biased" and some others have come under attack as being "fake news", efforts to make such labels stick have increased too. In some cases (e.g., InfoWars), the use of such labels is legitimate, because some online publishers deliberately spread conspiracy theories and false stories. Other news publishers are perceived as partisan and biased, in ways that damages their reporting credibility. Whether political bias affects journalism standards appears to be a debated topic with no clear consensus. Meanwhile, labels such as "far-left" or "alt-right" are highly contested and may become cause for prolonged edit wars on the Wikipedia pages of some news sources. In this paper, we try to shine a light into this phenomenon and its extent, in order to start a conversation within the Wikipedia community about transparent processes for assigning political orientation and journalistic reliability labels to news sources, especially to unfamiliar ones, which users would be more likely to verify by looking them up. As more of Wikipedia's content is used outside Wikipedia's "container" (e.g., in search results or by voice personal assistants), the issue of where certain statements appear in the Wikipedia page and their verifiability becomes an urgent one to consider not only by Wikipedia editors, but by third-party information providers too.

CCS CONCEPTS

• **Information systems → Wikis**; *Web search engines*; *Crowdsourcing*; • **Human-centered computing → Empirical studies in collaborative and social computing**; *Interactive systems and tools.*

KEYWORDS

wikipedia, google search, knowledge panels, political bias, editing

ACM Reference Format:
Khonzodakhon Umarova and Eni Mustafaraj. 2019. How Partisanship and Perceived Political Bias Affect Wikipedia Entries of News Sources. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316760

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316760

1 INTRODUCTION

Recently, a new kind of Wikipedia vandalism appeared in the news: the use of an extreme political label, in this case, "nazism", to describe the ideology of the California Republican Party. Interestingly, the public learned about it because Google showed this label in the Knowledge Panel of the California Republican Party's search engine result page (SERP). In fact, Google also received most of the blame for it,[1] despite the fact that the information was extracted from Wikipedia's page of this organization. Knowledge Panels (KP) in SERPs frequently contain information that comes directly from the Wikipedia article of the searched entity (a person, an organization, a company, etc.). Thus, edits in either the infobox[2] or the lead[3] of a Wikipedia entry may quickly find their way to Google's SERPs and therefore to a larger audience. In cases of blatant vandalism, such Wikipedia edits are reverted quickly (although Google search might need several hours to refresh its index that contains the changes). However, other labels describing the perceived political bias of a news publisher (rather than a self-declared political affiliation) or the lack of reliability of news sources are a matter of opinion and don't receive immediate resolution by Wikipedia editors, especially if they belong to pages that don't have a huge Wikipedia traffic. Such lack of clarity about how such labels are assigned, might have consequences in how these sources are viewed by the public.

One of our long-term research goals is to understand and support users' decision making about the credibility of online news sources. Through user studies [7, 10], we have identified that KPs in SERPs play an important role in this process, both through the amount of information they contain and its quality. Explicit references to perceived political bias of news sources, which are sometimes present in the Wikipedia lead of their articles are particularly helpful to users [10]. However, by comparing SERPs of news publishers over time,[4] we have noticed a few interesting patterns: for some news sources, the description swings between overly critic to fact-omitting (see case study in the next section); in others, the description relies on self-representation (citing the primary source) by burying down in the article secondary sources that contest the primary source's self-depiction.

Prior research [8] has established that when users read excerpts from Wikipedia content on Google search results, they often don't click-through to the entire article. Thus, these users might be at risk of creating a distorted or an incomplete impression of a news source based on a few sentences. Given the prominence of KPs

[1] https://www.politico.com/story/2018/05/31/mccarthy-slams-google-in-nazism-flap-616109
[2] A box of information that is displayed in most Wikipedia pages, on the right-side.
[3] The first paragraph in a Wikipedia page.
[4] https://medium.com/@enimust/the-information-panels-on-google-and-facebook-uncovering-their-blind-spots-2e8210b2e697

in SERPs (especially on mobile devices, where they are frequently placed ahead of organic search results, e.g., see Figure 2), the quality of excerpts from the Wikipedia leads, which are displayed in KPs, may potentially "make or break a reputation". Simultaneously, because of this importance, various actors might have an incentive to modify such content with the intention of positively or negatively impacting such reputation. Although issues related to the NPOV (neutral point of view) and notability are nothing new for Wikipedians, the increased political polarization in the society might pose new challenges that need to be addressed consistently.

In this preliminary work, we establish the grounds for studying such challenges systematically. Instead of considering broadly the issue of political bias and its detection in Wikipedia, as performed in [5], we make a deep-dive into a subset of Wikipedia articles of news sources for which we have scores of perceived political bias released by [9]. Initial results indicate that the "political labeling battle" through constant revisions is more fierce for news sources that are not very popular and might not have the same Wikipedia page protections in place as more popular news sources.

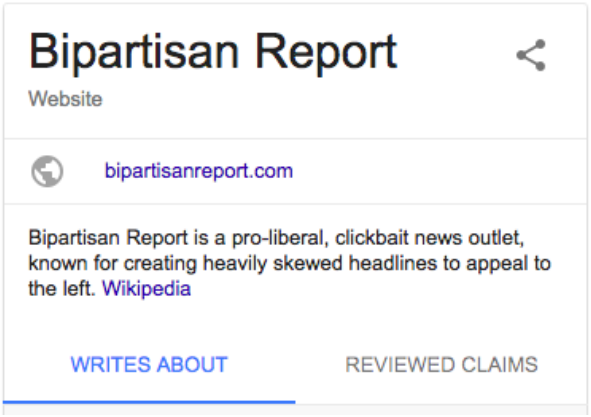


Figure 1: The Knowledge Panel for *Bipartisan Report* in January 2018. It contains a few sentences from the Wikipedia lead of this publication entry. Notice the mentions to perceived political bias and lack of reliability.

2 A CASE STUDY: BIPARTISAN REPORT

Research in news literacy has emphasized the use of "lateral reading" as a strategy for evaluating the reliability of news stories [13]. The first step of this strategy is to Google for the source of a news story. The KP containing information from Wikipedia is often prominent in Google's SERP for a query. One such example is shown in Figure 1, where the screenshot shows the KP for *Bipartisan Report* in January 2018. The brief description addresses this source's perceived political leaning (pro-liberal, left)[5] as well as lack of high journalistic standards (click-bait, heavily-skewed). When the page was created in December 2016, the description used stronger political labels:

> Bipartisan Report is a **far-left**, clickbait news outlet, known for creating heavily skewed headlines to appeal to the **far left**.

However, the author self-corrected, by replacing one instance of "far-left" with "pro-liberal" and removing "far" in the second "far left" expression, settling for the version shown in Figure 1. The complete text of the Wikipedia lead contained more details about this source's perceived lack of reliability, using a biased point of view,[6] quite similar to the language in Figure 1.

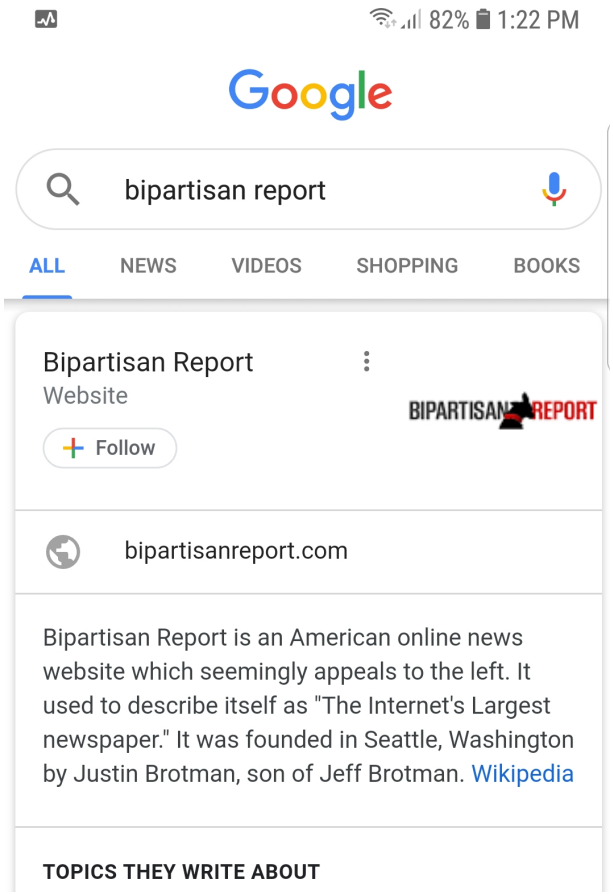


Figure 2: The Knowledge Panel for *Bipartisan Report* as of this writing (March 2019). Compared to the text in Figure 1) the Wikipedia lead was edited and now displays neutral language. Also, this screenshot was taken using a mobile device to indicate that the panel with Wikipedia content is very prominent in the search results shown in such devices.

It took 1.5 years, until May 2018, for the article's lead to be changed into a more neutral-sounding description, which is shown in Figure 2. As a result of a debate in the talk page, many contested parts of the article were removed too. The current version[7] is mostly neutral, but it also omits relevant details, such as several failed fact-checks of this source. Concretely, the fact-checker Snopes has

[5] No citations/references to support these political labels are provided in the Wikipedia article.

[6] https://en.wikipedia.org/w/index.php?title=Bipartisan_Report&oldid=753604666

[7] https://en.wikipedia.org/w/index.php?title=Bipartisan_Report&oldid=875424620

fact-checked four of *Bipartisan Report* stories, finding two of them False, one Mostly False, and one Unproven. The current Wikipedia article doesn't contain any of these facts about possible lack of reliability of *Bipartisan Report*.

If readers were to visit *Bipartisan Report's* Wikipedia page, they would see a big warning box listing multiple issues with the neutrality and quality of the article.[8] If furthermore, users were to visit the Talk page of the article, they would find a 4,600-word discussion, twenty times more text than to the 200-word article (including the warnings). However, few visitors come to this page. In 2018, it had a total of 8,769 page views and 11 edits.

There is an interesting conundrum here. Google continues to display the neutral-sounding description of *Bipartisan Report* in its KP, which doesn't contain any mention of the journalistic shortcomings of this news source, because Wikipedia editors couldn't agree on what to write. Meanwhile both Google and Wikipedia are providing a kind of legitimacy to this news source by "hiding" truly relevant information from users (its lack of reliability as a news source), while showing trivia information (the father of the founder happens to be a successful businessman).

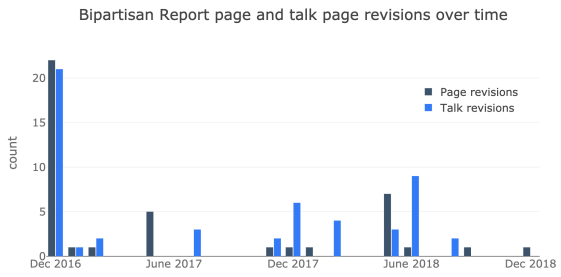


Figure 3: Comparison of edit distributions for the *Bipartisan Report* Wikipedia page and Talk page. Only discussions in May/June 2018 led to changes in the content of the page, by removing most of the contested information.

This case scenario raises many interesting questions worth discussing with respect to the design of information systems and their interfaces, as well as the ethical implications of such design choices:

(1) Why it took more than 1.5 years to take action about the lack of a neutral point of view? A warning label about how "The neutrality of this article is disputed" was added since the page creation in December 2016, and two more warnings about neutrality and bias were added in May 2017, but it took until May 2018 for any changes to happen.
(2) What did the editors discuss in the Talk page and why couldn't they agree on how to make the Wikipedia entry for *Bipartisan Report* more informative, especially by addressing issues of journalistic reliability? As Figure 3 shows, the talk page was active from the start, but that didn't initially affect the content of the page.
(3) Who should be responsible for misleading the public about this news source by omitting relevant information: Wikipedia editors who participated in the deliberation and couldn't agree, or Google that shows a KP based on Wikipedia content without any hints that this content is disputed within Wikipedia.

However, in order to motivate the need to answer such questions, we want to establish first the extent to which cases similar to *Bipartisan Report* occur in Wikipedia. After a discussion of partisanship and bias in the following section, we then explain our approach for discovering other cases potentially similar to *Bipartisan Report*.

3 PARTISANSHIP AND BIAS IN MEDIA

What makes *Bipartisan Report* an interesting study example is that it is perceived by certain readers as a partisan outlet, politically biased to entirely favor one side. The issue of partisanship in media is one with a long history. In the United States, partisan media was the only kind of media in existence for a long time [11]. However, over many decades, newspapers slowly freed themselves from political parties' influence and created a "norm of objectivity" [12] to focus on reporting the facts. Interestingly, this process happened mostly in the United States, and as a result, established newspapers do not identify as having a political orientation (such as left of right). Meanwhile, such labels are very common in Europe, where newspapers continue to be openly aligned with certain political ideologies. In fact, Wikipedia itself is a good place to experience this distinction. While the Wikipedia pages for the list of newspapers in Germany,[9] or the UK[10] contain an explicit column for the political alignment of newspapers, the corresponding page for the United States[11] contains no such column.

However, with the advent of Cable TV, conservative radio, and then later the web, the media landscape has changed in the United States too. Conservative elites (for example, radio and TV hosts such as Rush Limbaugh and Bill O'Reilly) have for years argued that the media in general has a "liberal bias" [3], in order to justify the need for conservative outlets. Researchers have argued that there is even a "Limbaugh effect" that could explain increased polarization in the American public [6]. In such a polarized environment, even though the majority of news sources in the United States remain objective in their reporting, the public perceives them increasingly as biased, as polls have been showing over the years.[12]. Moreover, the establishment of openly partisan web-based outlets has further contributed in the creation of a media ecosystem in which users feel skeptic and want to ascertain the bias of every news publisher.

Researchers have been investigating various methods to characterize the possible bias (or slant) of various news publishers. One method relies on the similarity of language between politicians and newspapers [4], another on a combination of crowdsourcing and machine learning decision-making [2]. In contrast to these content-based approaches, we also find in the literature approaches that

[8] However, this is not true if one visits the Wikipedia page on a mobile device. On it, the warning box is not visible, it's a link to click, further hiding the information about the quality of the page.

[9] https://en.wikipedia.org/wiki/List_of_newspapers_in_Germany
[10] https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_Kingdom
[11] https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States
[12] https://news.gallup.com/poll/225755/americans-news-bias-name-neutral-source.aspx

rely on the ideological alignment of audiences [1]. We rely in one such approach to find news publishers with a score of perceived bias.

4 DATA COLLECTION: SOURCES WITH PERCEIVED BIAS

A common limitation of the content-based approaches mentioned above is the limited size of the news sources they rate. For example, [2] focuses only on 15 news sources. In comparison, the audience-based approach that works by "polling" the users who share news stories in social networks such as Facebook and Twitter have the potential to scale and provide information for a large number of news sources. We explain such an existing approach in the following and then make use of its dataset to discover news sources with Wikipedia pages for our analysis.

4.1 Partisan Audience Bias

In [9], the authors created the Partisan Audience Bias (PAB) dataset[13] from links shared by real users on Twitter. Using voter registration records of US citizens with republican and democratic affiliations, the study identified 519,000 Twitter accounts matching these citizens. Over a certain period of time, a set of 113 million tweets by these accounts were collected and only tweets with URLs were kept. The URLs were processed to extract the second-level domain names. For example, `http://www.bbc.com/news/business-38686568` was converted into `bbc.com`. Then, to reduce noise, since 63% of links were shared only once, the authors kept only the domains that were shared more than 50 times (by different users). This led to a dataset of 19,022 sites. For each site, the authors calculated a bias score between -1 (a site shared only by democratic voters) to +1 (a site shared only by republican voters). Sites that get a bias score between -1 and +1 were shared by a mix of democratic and republican voters. For example, *The Wall Street Journal* had a score of 0.0106, signaling that it is a news source shared almost equally by both sides. Meanwhile, the bias score for *Bipartisan Report* was -0.7469, and the score for *InfoWars* was 0.7817, putting these two sources at opposite ends. However, the study's authors caution the readers that these scores are not absolute but only relative, since there is no fixed zero point in assigning bias.

4.2 SERP pages for the PAB dataset

Using one computer with a Chrome browser in incognito mode, and creating one new instance of the browser for each query, we automatically collected the Google SERPs for all 19,022 site names in the PAB dataset. By parsing the SERPs, we found 3,763 SERPs (or 19.8%) that contained a KP. 89.3% of these KP contained a Wikipedia link, indicating their provenance from Wikipedia. Given that the use of URLs as search queries doesn't guarantee good search results, we decided to rectify this issue with a second collection. We automatically performed approximate string matching of the URLs from the PAB dataset with the list of URLs and their titles from Google's SERP in order to find textual phrases to search. For example, the URL `aflcio.org` from the PAB dataset is matched with three pairs: (`aflcio.org`, *AFL-CIO: America's Unions*), (`wyomingaflcio.org`, *Wyoming State AFL-CIO*), and (`www.massaflcio.org`, *Massachusetts AFL-CIO*). Using this technique, we were able to find matches for 17,497 URLs. We repeated the SERP collection and this time found 3,539 KPs (19.3%), again with the majority with Wikipedia links. At the end of this two-step SERP collection, we had a list of 4,967 matches between (URL, score) pairs from the PAB dataset and Wikipedia links that had appeared in KPs for our two data collections.

4.3 Two lists of online news sites

Inspecting the created dataset of Wikipedia links and KPs description, we noticed that many sites are not news sources. As expected, people share on Twitter other links too, such as event descriptions, products, etc. Since our focus is on news sources, we selected a subset of 1,391 URLs that come from the two categories below:

- *Newspaper news sites* - a list that contains all sources, whose KP has a "Newspaper" category. It is made of 695 unique sites.
- *Non-newspaper news sites* - a list that contains all sources, whose KP short description contains the word "news". There are 696 unique sites in this list.

We acknowledge that other links in our dataset might be news sources, and identifying them remains for future work.

4.4 Collecting Wikipedia Leads

We are interested in how news sources are labeled in terms of their perceived bias and reliability and such labeling often happens in the lead paragraph(s) of the Wikipedia entry. Therefore, our current focus is on examining how Wikipedia leads change over periods of time. To that purpose, we collected all revisions of pages corresponding to our news sources dataset. The dataset was collected on 01/12/2019, and hence contains revisions between February 2001 and January 2019. Deleted or inaccessible revisions are excluded. Overall, for 1,391 news-related sources from the two lists, our dataset contains 434,923 revisions, 100,269 of which occur in the lead section.

Using the MediaWiki action API[14], we collected all revisions for each page, including the revision id, time-stamp, user who made the revision, and comments. For every revision, the lead section (or everything that comes before the first section of Wikipedia article) was also extracted. After filtering out Wiki code, such as templates, style tags, and references, we obtained the plain-text of the lead section.

5 RESULTS

In a political atmosphere of polarization, individuals at opposite sides are applying political labels to media outlets, with the purpose of devaluing their reporting (by implicitly considering them as biased and unfair). For example, President Trump tweeted in August 2018 the following: "Republican/Conservative & Fair media is shut out. Illegal? 96% of results on "Trump News" is from National Left-Wing Media"[15]. To examine whether this process of political labeling is happening in Wikipedia too, we compiled a list

[13] Available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QAN5VX

[14] https://www.mediawiki.org/wiki/API:Main_page

[15] https://www.politifact.com/truth-o-meter/statements/2018/aug/29/donald-trump/no-96-google-news-stories-trump-arent-left-wing-ou/

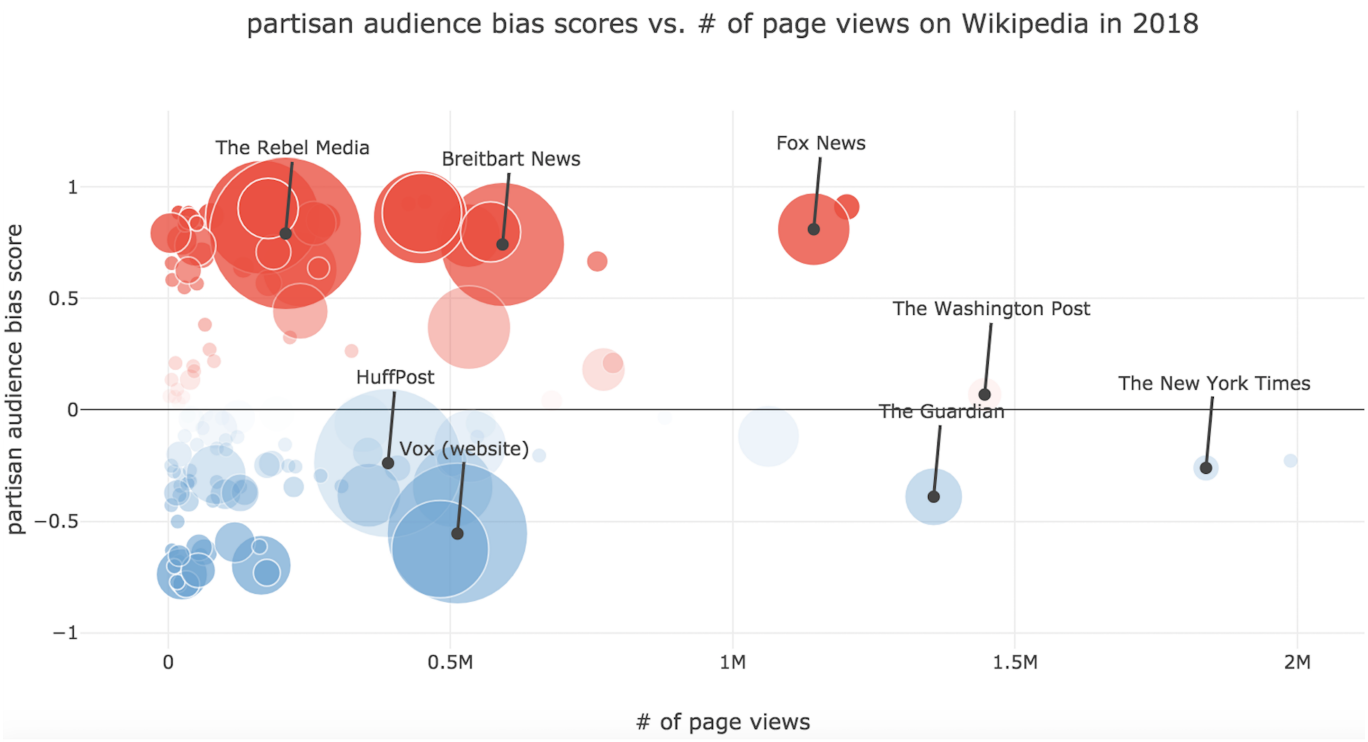


Figure 4: A visual representation of the "political labeling battles" in the Wikipedia pages of news sources that are perceived as politically biased. For visual clarity purposes, we plotted a subset of 127 sources, for which political labels were removed/re-added more than 5 times in 2018. Due to this decision, many sources with perceived bias scores between -0.5 and 0.5 are not displayed, because they didn't pass the threshold.

Table 1: Number of Wikipedia pages for news sources that contain a political label in their latest revision.

Left-leaning	left (46), democrat (6), progressive (18), liberal (35)
Right-leaning	right (49), republican (1), conservative (51)

of target words that are associated with opposite political leanings: 'left' and 'right'; 'democratic' and 'republican', as well as 'liberal', 'progressive' and 'conservative'. By inspecting changes in the lead, we identify revisions in which these words were repeatedly added in or removed from the Wikipedia lead section, an action that indicates lack of consensus about the use of such political labels. For the latest revisions of Wikipedia articles in our dataset, we tabulate the presence of such labels in Table 1. We can notice a somewhat balanced use, 101 are right-leaning labels and 106 are left-leaning. Meanwhile, the total number is small (207) compared to the list of 1,391 sources, indicating that for the majority of news sources, no such political labels are used.

5.1 Changes Related To Political Bias

When Wikipedia editors don't agree about the use of a political label for a news source, they might engage in an edit war to either preserve or remove a label. Are such edit wars correlated with the perceived bias? Although in this paper we do not formally test such an hypothesis, we begin its exploration through the visualization in Figure 4, which illustrates the relationship between (1) the partisan audience bias score of a news source (between -1 and +1, shown in the y-axis), (2) the number of views its English Wikipedia page received in 2018 (shown in the x-axis), and the number of times the political labels from Table 1 were repeatedly removed and added back to the article lead in 2018. The third variable is expressed through the radius of markers.

There are several interesting trends we observe here. On the farther end of x-axis, we have popular websites, such as *Fox News* and *The New York Times*, whose Wikipedia pages received over 1 million views in 2018. These pages have overall a larger number of page revisions (339 and 183 respectively) more than other pages in our list (89.1 in average), however, such revisions are rarely about the political labels in question, and thus these sources appear with a small radius in the graph. Meanwhile, somewhat less popular pages, such as *HuffPost* or *The Rebel Media* that received in the range of 500,000 views, had a large number of changes in political labels, contributing to the depiction in the graph with a larger radius. Such changes might be due to disagreements on the extent to which a certain political label could be applied to these news sources. It is however important to note that both *Fox News* and *The New*

York Times pages are under Wikipedia's semi-protection policy, as opposed to *HuffPost* and *The Rebel Media*, which can be edited by unregistered and newly registered users.

Most importantly, Figure 4 suggests that news sources that are perceived as strongly biased (on both sides of the political spectrum) are targets for these political labelling actions, even though many of these news sources are not popular in terms of page views.

5.2 User participation in revisions

In addition to investigating the nature of edits in the Wikipedia lead section, we were also interested in identifying the involved users. For each revision that was associated with addition and removal of political labels, we identified its editor and when available collected information about their account, such as the registration date, total number of contributions, and permissions using the Media Wiki action API.

In our analysis, we considered two groups of editors: (1) users who added a political label at some point in the lead of a page, and (2) users who removed a political label. Overall there are 3,946 and 3,796 unique users in each group respectively, including unregistered editors (identified by their IP address), and 1,368 users who occur in both groups.

We found that for most editors from both groups, their scope of edits related to political labelling in the lead is focused on a limited number of pages. In fact, about 90-91% of users from each group edit only one Wikipedia page from our lists of newspapers and news-related sources. There are a few users whose edits span over as many as 30-31 different pages. These are primarily Wikipedia anti-vandalism bots.

Lastly, we analyzed the age of accounts of users who added/removed a political label at the time the revision occurred. Figure 5 shows that most revisions related to political labelling were made by accounts which are 1 to 2 years old. This can be observed in both groups of editors, though "younger" accounts are involved in adding such labels (notice the visible blue spike). We also couldn't find a statistically significant difference between the distributions.

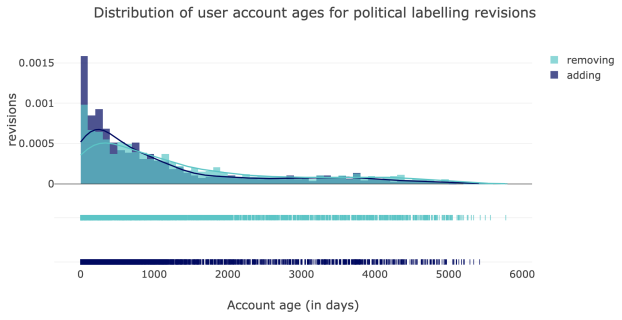


Figure 5: Comparison of distribution of ages for accounts that make edits involving addition/removal of political labels in the Wikipedia article lead.

6 CONCLUSION

Motivated by the presence of Wikipedia content in third-party platforms, such as Google's search results, we seek to identify the extent to which the characterization of news publishers is affected by edit wars in Wikipedia, in the backdrop of a polarized society. Our initial results, which are limited to mostly North American news sources, indicate that many pages that don't have editing protections in place are affected by continuous efforts to add/remove political bias labels to their lead text. That is, this problem is not unique to our case study, *Bipartisan Report*, therefore, it merits further research. Inspecting the editors involved in such edits, we found relatively sizeable groups on both sides of the edit wars, indicating that this is not an isolated issue, and that it should be considered for discussion in the Wikipedia governance deliberation processes. This research has also implications for web literacy efforts. Learning to recognize warning labels, consulting talk pages, and looking at revision histories of Wikipedia pages needs to be one of the skills that is explicitly taught as part of news literacy programs.

ACKNOWLEDGMENTS

The authors are grateful to the members of the Wellesley Cred Lab for their support. This work is partially supported by the National Science Foundation through grant IIS 1751087.

REFERENCES

[1] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132. https://doi.org/10.1126/science.aaa1160 arXiv:http://science.sciencemag.org/content/348/6239/1130.full.pdf

[2] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.

[3] David Domke, Mark D Watts, Dhavan V Shah, and David P Fan. 1999. The politics of conservative elites and the âĂIJliberal mediaâĂİ argument. *Journal of Communication* 49, 4 (1999), 35–58.

[4] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 1 (2010), 35–71.

[5] Christoph Hube and Besnik Fetahu. 2018. Detecting Biased Statements in Wikipedia. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1779–1786. https://doi.org/10.1145/3184558.3191640

[6] David A Jones. 2002. The polarizing effect of new media messages. *International Journal of Public Opinion Research* 14, 2 (2002), 158–174.

[7] Emma Lurie and Eni Mustafaraj. 2018. Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources. In *Web Science*. ACM, 107–116.

[8] Connor McMahon, Isaac L Johnson, and Brent J Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies.. In *ICWSM*. 142–151.

[9] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias Within Google Search. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 148 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274417

[10] Annabel Rothshild, Emma Lurie, and Eni Mustafaraj. 2019. How the Interplay of Google and Wikipedia Affects Perceptions of Online News Sources. In *Computation + Journalism Symposium*.

[11] Michael Schudson. 1978. *Discovering the news: A social history of American newspapers*. Basic Books.

[12] Michael Schudson. 2001. The objectivity norm in American journalism. *Journalism* 2, 2 (2001), 149–170.

[13] Sam Wineburg and Sarah McGrew. 2017. Lateral Reading: Reading Less and Learning More When Evaluating Digital Information. *Stanford History Education Group Working Paper No. 2017-A1* (2017). https://ssrn.com/abstract=3048994

Detecting and Gauging Impact on Wikipedia Page Views

Xiaoxi Chelsy Xie
cxie@wikimedia.org
Wikimedia Foundation

Isaac Johnson
isaac@wikimedia.org
Wikimedia Foundation

Anne Gomez
agomez@wikimedia.org
Wikimedia Foundation

ABSTRACT

Understanding how various external campaigns or events affect readership on Wikipedia is important to efforts aimed at improving awareness and access to its content. In this paper, we consider how to build time-series models aimed at predicting page views on Wikipedia with the goal of detecting whether there are significant changes to the existing trends. We test these models on two different events: a video campaign aimed at increasing awareness of Hindi Wikipedia in India and the page preview feature roll-out—a means of accessing Wikipedia content without actually visiting the pages—on English and German Wikipedia. Our models effectively estimate the impact of page preview roll-out, but do not detect a significant change following the video campaign in India. We also discuss the utility of other geographies or language editions for predicting page views from a given area on a given language edition.

CCS CONCEPTS

• **Information systems** → **Web mining**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*.

KEYWORDS

wikipedia; bayesian structural time series; page views; causal inference

ACM Reference Format:
Xiaoxi Chelsy Xie, Isaac Johnson, and Anne Gomez. 2019. Detecting and Gauging Impact on Wikipedia Page Views. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316751

1 INTRODUCTION

Wikipedia is the fifth-most-visited website worldwide [1] at 190 billion page views in 2018 alone [7] and is turned to by readers for all sorts of reasons ranging from simple curiosity to fact-checking to making a personal decision [20]. Despite its success, there are still many regions in the world where it is relatively unknown [9] or access is blocked [6]. In an attempt to improve access and awareness worldwide, the Wikimedia Foundation (WMF) has conducted various campaigns and efforts aimed at improving access to Wikipedia in various regions.[1]

[1]https://meta.wikimedia.org/wiki/New_Readers

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316751

Many researchers have sought to estimate the impact of external events on Wikipedia page views. This has included the effect of posting Wikipedia articles on external websites [14, 23], privacy concerns on viewing of sensitive Wikipedia articles [15], and censorship [24]. Conversely, much research has also sought to use Wikipedia page views as a predictor—i.e. *nowcasting* or *forecasting* [16]—of external entities such as the stock market [13], box office returns for movies [12], and disease incidence [16].

Evaluating the impact, however, of a given campaign or external event can be difficult. Daily page views to Wikipedia projects can be quite noisy, being affected by weekly, seasonal, holiday-related trends [21]. A change in the volume of page views following an awareness campaign could also easily be the result of an unrelated event—e.g., a celebrity marriage or World Cup game [7]. To account for these challenges, many researchers rely on some form of regression discontinuity design that focuses on changes between a short time period (e.g., two weeks) before and after an event (e.g., [14, 23, 24]). While powerful, this approach is limited to studying short-term effects and does not naturally lend itself to the task of nowcasting or forecasting.

In this paper, we explore the utility of Bayesian structural time series (BSTS) models for evaluating the impact of external events. BSTS models are designed to predict a given time series based on historical data, seasonality components, and additional control time series. They naturally incorporate uncertainty and the resulting forecast can then serve as a counterfactual—i.e. be compared against the actual time series following a given intervention to determine whether there is evidence that the intervention increased or decreased the magnitude of the time series. We test the BSTS model in two scenarios: the page preview roll-out in German and English Wikipedia as well as a video campaign in India designed to increase awareness about Hindi Wikipedia.

Our contributions are as follows:

- **Page Previews**: using the roll-out of page previews on the German and English Wikipedia, we demonstrate that our BSTS model can effectively detect changes in page views given predictive control series.
- **Impact of video campaigns in India**: applying our BSTS model to online and TV awareness campaigns in India, we do not find evidence of increased page views as a result of the online or TV campaigns.
- **Correlations across languages and regions**: we evaluate the predictive power of page-view time-series between pairs of Wikipedia language editions and regions. We find evidence that page view trends are unique to a given country and language edition and that control series ideally originate from the same language edition followed by same country to be a useful predictor.

2 RELATED WORK

In this work, we draw methods from the time series prediction literature and motivation from the literature that has sought to understand the impact of external events on Wikipedia activity.

2.1 Time Series Modeling

The goal of the time series modeling that we employ in this paper is to understand whether a specific intervention significantly impacts a given metric for which we have temporal data—e.g., whether the roll-out of a new feature causes a change in daily traffic. There are many approaches to time series modeling that span from quite simple to much more complex in accordance with how many assumptions they make. Common to these models, however, is that their validity depends on the model being able to make direct comparisons between the time series prior to an intervention and the time series following the intervention [2]. Threats to this validity come from a variety of sources that may affect time series independent of the intervention being studied: seasonality effects such as natural variation by day of week or month of year, unaccounted external events such as holidays or changes in the size of the underlying population. The likelihood that these events affect the time series increases as the time period being studied increases. A strong model, then, incorporates covariates that can control for seasonality, holidays, and other external factors that might affect the time series. A strong model also effectively represents its own uncertainty about predictions when there are insufficiently strong controls in place.

The core distinguishing features of approaches are 1) whether they include a control time series, and, 2) whether they directly compare metrics before and after the intervention or predict the time series after the intervention and compare this counterfactual prediction with the actual data. A control time series is a time series that is highly correlated with the "treated" times series, but, importantly, is known to not be affected by the intervention. The value of a control time series is that it helps to ensure that if seasonality or another event affects the treated time series, this effect is not conflated with the impact of the intervention because the effect should also be present in the control time series. The value of comparing the post-intervention time series with a counterfactual, as opposed to just the values from prior to the intervention, is greater flexibility to changing conditions. A model that produces the counterfactual can take into account more data about the time series prior to the intervention and, therefore, better account for shifts in covariates that might occur following the intervention. This is especially important when considering the long-term effects of an intervention. For these models, if the actual time series falls outside of the bounds of the counterfactual time series, this provides evidence that the intervention had a significant impact.

There are many considerations for how to build a robust time series model in order to produce the counterfactual predictions. Primarily, some models that use static regression to produce the counterfactual predictions falsely assume the data to be independent and identically distributed (i.i.d), which would result in an underestimation of the uncertainty [3]. Secondarily, to help avoid over-fitting, we need to choose appropriate control time series. Castle et al. [5] reviews and compares 21 methods for variable selection, including significance testing (e.g., forward and backward step-wise regression) and information criteria (e.g., AIC, BIC). Other popular model selection algorithms in time-series forecasting includes principal component and factor models, and penalized regression models (e.g., Lasso, ridge regression). However, these techniques force us to use a fixed set of selected variables, or do not account for the uncertainty in variable selection. Lastly, in order to gauge the uncertainty of the impact, we need to account for various sources of uncertainty in the model. Besides the uncertainty in variable selection and auto-correlation mentioned above, we also want to account for uncertainties in the historical relationships between treated and control time series, as well as uncertainties in seasonality and other components in the model.

2.2 Impact of External Events on Wikipedia

A number of papers have considered the challenge of establishing how an external event has affected dynamics within Wikipedia. Vincent et al. [23] and Moyer et al. [14] take what is known as a *interrupted time series* (ITS) approach to examine how posts on Reddit that include links to Wikipedia articles affect page view traffic on Wikipedia. Both model a Reddit post with a Wikipedia link as a "shock" to that Wikipedia article and compare the mean number of page views in a short period of time before and after the post to determine whether there is a significant difference. Zhang and Zhu [24] take a similar approach for the rate of contributions to Chinese Wikipedia from outside editors before and after a block on mainland China. Zhang and Zhu also seek to control for seasonal trends by examining the same time period in prior years. These approaches build on the assumption that there should be no difference in the expected page views between the pre- and post-intervention periods, and therefore any difference in page views can be causally tied to the Reddit post or block. As discussed above, these are assumptions that may hold true for short time-spans like one week, given that they do not happen to coincide with major holidays or events. This assumption is increasingly tenuous, however, as more long-term trends are considered.

Penney [15] also starts with an ITS model to understand the impact of the Edward Snowden revelations on page views to "terrorism-related" Wikipedia articles. Notably, because Penney examines a much longer time-period comprising 32 months, their analysis also includes a "control time series" that are security-related articles that are similar in content but less likely to be affected by the Snowden revelations. This approach is often referred to as difference-in-differences (DD) and is similar to how we construct the BSTS models considered in this work, but the BSTS models directly incorporate the concept of a control series as a core component of the models and captures the uncertainty of the relationship between the treated series and the control series. This makes for a much more explicit and robust means of controlling for additional external effects that may otherwise be conflated with the treatment under study.

3 METHODS

We use a single time series model architecture, described below, and apply it to two different events. Each event involves an external

event that led to a potential shift in page views. We describe each event alongside its results.

3.1 Bayesian Structural Time Series Model

In this work, we use Bayesian structural time series (BSTS) model [19]. Per the components discussed in §2.1, these models can incorporate control covariates and time series, generate counterfactual predictions of page views for the post-intervention period assuming that the intervention did not take place, and naturally model their own uncertainty about these counterfactuals. We can then compare the counterfactual predictions and actual page views to quantify the causal impact of the intervention.

BSTS models combine three statistical methods into an integrated architecture [18]:

- A structural time series model for trend and seasonality, estimated using Kalman filters;
- Spike and slab regression for variable selection;
- Bayesian model averaging for the final prediction.

Structural Time Series Model: Under different assumptions, a very large class of models can be expressed in the form of structural time series models, including all ARIMA models [10]. This flexibility allows BSTS models to accommodate multiple sources of variations, including trends, seasonality, and latent evolutions of the treated series that cannot be explained by known trends or events. Specifically, a structural time series model (e.g. Eq. 1) decomposes the time series into four components: a level (μ_t), a local trend or slope (δ_t), seasonal effects (τ_t) and error terms. The model described here adds a regression component ($\beta^T \mathbf{x}_t$) to incorporate the control time series and other covariates. It is a stochastic generalization of the constant-trend regression model (e.g. $y_t = \mu + \delta t + \beta^T \mathbf{x}_t + \epsilon_t$), where the level μ_t and slope δ_t parameters each follow a random walk model instead of a constant. This allows for greater flexibility in the trends expressed within the model.

$$
\begin{aligned}
y_t &= \mu_t + \tau_t + \beta^T \mathbf{x}_t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2) \\
\mu_t &= \mu_{t-1} + \delta_{t-1} + u_t, u_t \sim N(0, \sigma_u^2) \\
\delta_t &= \delta_{t-1} + v_t, v_t \sim N(0, \sigma_v^2) \\
\tau_t &= -\sum_{s=1}^{S-1} \tau_{t-s} + w_t, w_t \sim N(0, \sigma_w^2)
\end{aligned}
\tag{1}
$$

Spike and Slab: There are often many potential control series but including them all would likely lead to over-fitting and very complex models. A spike-and-slab prior over coefficients [8, 11] is designed to solve this challenge. The spike part controls the probability of whether a given variable would be chosen for the model—i.e. having a non-zero coefficient. The slab part shrinks the non-zero coefficients toward prior expectations which is often zero. Upon observing data, Bayes' theorem updates the inclusion probability of each coefficient. Then when sampling from the posterior distribution of a regression model, many of the simulated regression coefficients will be exactly zero [17].

Bayesian Model Averaging: To generate counterfactual predictions, the procedure uses the Markov chain Monte Carlo (MCMC) algorithm to draw samples from the parameter's posterior distribution and then combine that with the available data to yield a distribution of the counterfactual predictions. The model can then compute the difference between the actual values of a treated series in the post-intervention period and the distribution of counterfactual samples to yield an estimate of the distribution of the impact [4]. Because the structural time series model, spike-and-slab regression and model averaging all have natural Bayesian interpretations, BSTS is able to account for various sources of uncertainties using MCMC. This allows us to gauge confidence in the magnitude of causal impact and estimate the posterior probability that the causal impact is non-existent.

We use the *BSTS*[2] and *CausalImpact*[3] R packages to fit the BSTS models. The following parameters comprise each BSTS model. Figure 1 illustrates the parameters of the model for Hindi online video campaign in §4.2.

- **Treated Time Series**: this is the time series under study—e.g., daily internally-referred page views to English Wikipedia over the period of several months.
- **Intervention**: an event that occurred on a specific date during the study period that is believed to have affected the treated time series—e.g., the introduction of a new feature onto Wikipedia that might change the daily number of page views. Model validation is conducted entirely on data prior to the intervention.
- **Pre-Intervention Period**: time period from the first data point of the treated series to the day before the intervention. For each model in this work, we explore four different pre-intervention period length using grid-search: 12 weeks, 18 weeks, 183 days and 400 days.
- **Post-Intervention Period**: time period following the intervention for which the impact is being estimated—e.g., daily page views for the six weeks following the roll-out of a new feature. For each model in this work, we set this to 6 weeks.
- **Covariates**: additional variables that help explain the treated time series—e.g., total population online in a country. This also includes the control series described below.
- **Control Series**: a time series that is predictive of the treated time series prior to the intervention, but that is not be impacted by the intervention—e.g., daily page views for a similar Wikipedia edition for which the feature was not rolled out. The authors of the *CausalImpact* library we use for estimating the models suggest using 3-50 covariates.[4] Thus, for some models in this work where we have hundreds of control series—e.g., many combinations of different regions and language editions—we use correlation and dynamic time warping (DTW) [22] algorithms with pre-intervention data to prescreen and trim the list of control series before feeding them into the BSTS model.
- **Seasonality**: weekly and seasonal trends, or holiday effects that did not get captured by the control series. For each model in this work, we include features for day-of-week and month. We also include major holidays for the regions under study as described in the Results section.

[2] https://cran.r-project.org/package=bsts
[3] https://cran.r-project.org/package=CausalImpact
[4] https://stats.stackexchange.com/questions/162930/causalimpact-should-i-use-more-than-one-control/163554#163554

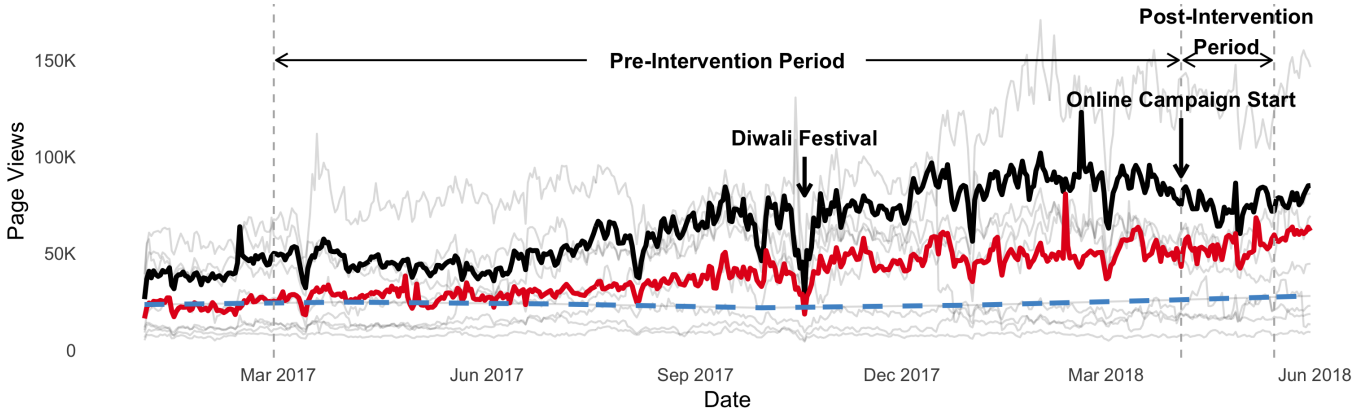


Figure 1: Hindi Wikipedia daily externally-referred page views from the top 10 states with the most page views in India. The black (top highlighted) series is the treated series: page views from Madhya Pradesh. The red (bottom highlighted) series is one of the most predictive control series according to the model in §4.2: page views from the neighboring state of Rajasthan. The blue dashed line is a covariate: interpolated number of internet subscribers in thousands from Madhya Pradesh. The intervention under study (online video campaign) started on 3 April 2018.

- **Trend Model**: the architecture for the model that predicts how the treated time series evolves. For each model in this work, we explore the following types of trend components using grid-search: local level, local linear, semi-local linear, and static intercept term.

4 RESULTS

Below we describe the context for two events on which we tested our BSTS model and their results.

4.1 Wikipedia Page Previews

Beginning in 2014, Wikipedia began exploring a new feature that would allow for page previews for the desktop version of the site. When a user moused over a link, a card would appear with an image and part of the first paragraph from the article that the link pointed to (see the Wikimedia Blog post[5] for more information and an example of a page preview on Wikimedia Commons[6]). This allows users to preview the article content without clicking on it (and thereby recording a page view). As a result of this ability to explore Wikipedia content without actually visiting the pages, it was expected that page views would actually drop with the roll-out of this feature.

In late 2017, the feature was finally rolled out to a proportion of anonymous users on German (de-wiki) and English (en-wiki) Wikipedia in a series of A/B tests. These tests were analyzed and it was determined that page preview feature led to a drop of approximately 4% in page views across these Wikipedia communities.[7] The full deployment of page previews to all anonymous users of these communities occurred on April 11 and 17 respectively. These A/B tests present us with an opportunity to explore the power of our BSTS models because they experimentally determined the expected effect on overall page views from the full deployment of page previews. Specifically, from the A/B tests starting on December 21 2017, in which 1.5% and 4% of anonymous users on English and German Wikipedias by default had access to page preview functionality respectively, we expect our BSTS model to detect:

- **de-wiki**: a 3.0% decrease in page views after April 11 2018.[8]
- **en-wiki**: a 4.7% decrease in page views after April 17 2018.[9]

4.1.1 Model Parameters. For each model, we set the pre-intervention period to be 400 days and the post-intervention period to be 6 weeks. For English Wikipedia, this means that the time series starts on 13 March 2017 and includes daily page views data through 28 May 2018, with the intervention occurring on 17 April 2018. For German Wikipedia, this means that the time series starts on 7 March 2017 and includes daily page views data through 22 May 2018, with the intervention occurring on 11 April 2018. Alongside day-of-week and monthly seasonality, we also include the following holidays: Christmas and New Year's. For trend modeling, we choose a static intercept term for German Wikipedia—i.e. we expect the trend of the time series to be soaked up by the regression component, and a local level model for English Wikipedia—i.e. the trend will be predicted around the weighted average values of recent observations.

For the control series, we rely on the assumption that while page previews should impact the internally-referred page views—i.e. page views as a result of navigating from one Wikipedia page to another—there is no reason that the previews would impact externally-referred page views—i.e. page views that result from someone navigating from a search engine or other, non-Wikimedia website—or direct page views without referrer. Specifically, we select the daily internally-referred page views from en-wiki or de-wiki as the treated time series. For our control time series, we include the daily externally-referred page views and direct page views from the same Wikipedia language edition as the treated time series under

[5]https://blog.wikimedia.org/2018/05/09/page-previews-documentation/
[6]https://commons.wikimedia.org/w/index.php?curid=47213242
[7]https://www.mediawiki.org/wiki/Page_Previews/2017-18_A/B_Tests
[8]https://phabricator.wikimedia.org/T191966
[9]https://phabricator.wikimedia.org/T191101

study (i.e. en-wiki or de-wiki), as well as daily externally-referred page views from the other top-20 largest Wikipedia editions (e.g. Russian and Spanish Wikipedia).

4.1.2 Results. We find that our time series models for both English and German Wikipedia are quite accurate. The validation statistics associated with the model provide an indication of how effective the model was at predicting the pre-intervention time series. With 10-fold cross validation and prediction evaluated on 6 weeks of daily page views (from the end of the pre-intervention period), the holdout mean absolute percentage error (MAPE) of the English Wikipedia model is 2.54%, and the holdout MAPE of the German Wikipedia model is 3.92%.

Turning to the estimate of the causal impact of the page preview roll-out, the treated time series along with counterfactual estimates from model for en-wiki are shown in Figure 2. Recall that the early A/B tests indicated that there would be a 4.7% decrease in page views for en-wiki. Our BSTS model, using the externally-referred and direct page views as control series, estimates a 3.0% decrease and correctly determines that no impact—i.e. 0% change—falls outside of the 95% credible interval [1.9%, 3.9%], indicating that the roll-out resulted in a significant change in page views. The most predictive control series in this model is the search engine referred page views on English Wikipedia with an average standardized coefficient of 0.65—i.e. when search engine referred page views change 1 standard deviation, we expect to see internally-referred page views change 0.65 standard deviation, and the posterior inclusion probability—i.e. the probability of this coefficient being different from zero—is 100%.

We see analogous results for de-wiki: the BSTS model estimated a 2.6% decrease in page views with a 95% credible interval of [1.9%, 3.4%], in line with the 3.0% decrease that had been determined via A/B testing. Similarly, the most predictive control series in this model is the search engine referred page views on German Wikipedia with an average standardized coefficient of 0.95 and a posterior inclusion probability of 100%.

4.2 Hindi Video Campaign

In India, only 33% of Hindi internet users have heard of Wikipedia [9] and, while there are 120,000 Wikipedia articles in Hindi, many people do not know that Hindi content exists. Meanwhile, internet access is growing 20%+ per year across India[10] and Hindi online content consumption is growing 94% per year[11]. In July 2017, the Wikimedia Foundation and the Hindi Wikimedians User Group began collaborating to reach "New Readers" in India through production and promotion of an online video.[12] The goal is to increase awareness and drive new usage of Wikipedia among Hindi speaking internet users.

To explain and promote Hindi Wikipedia (hi-wiki), the Wikimedia Foundation launched the video campaign on 3 April 2018. The Ektara[13] video was promoted on YouTube and Facebook targeting Hindi internet users in Madhya Pradesh, many of whom who had not heard of Wikipedia. The online promotion ran for three weeks and the video gathered 2.61 million views. This was followed by a second push over TV during a major Cricket event (on DD Sports during the Indian Premier League finals) on 27 May 2018 to the whole country, which reached 1.37 million viewers.[14]

4.2.1 Model Parameters. The pre-intervention period is 400 days and the post-intervention period is 6 weeks. Alongside day-of-week and monthly seasonality, we also include the following major Hindu holidays: Diwali, Raksha Bandhan, Holi, Dussehra, and New Year.[15] Local level model and a static intercept term are chosen as the trend for online campaign model and TV campaign model respectively.

For the evaluation of the impact of the online campaign, we set the treated time series to be daily externally-referred page views to hi-wiki from the Indian state of Madhya Pradesh (as determined by IP geolocation) because the promotion of the online video campaign was targeted at Madhya Pradesh. We select the externally-referred page views because it is a good indicator of the general brand awareness. The time series starts on 27 February 2017 and includes daily page view data through 14 May 2018, where the intervention occurred on 3 April 2018. For the control series, we use daily hi-wiki page views, as well as page views to other popular Wikipedia language editions and Wikimedia projects,[16] from the rest of India by states. We also included the daily number of internet subscribers in Madhya Pradesh as a covariate, which is linearly interpolated from a quarterly series reported by Telecom Regulatory Authority of India.[17]

For the evaluation of the impact of the TV campaign, we set the treated time series to be daily externally-referred hi-wiki page views from the entire country of India because there was no state-specific targeting of the campaign. The time series starts on 22 April 2017 and includes daily page view data through 7 July 2018, where the intervention occurred on 27 May 2018. For the control series, we use daily hi-wiki page views and page views to other popular Wikipedia language editions and Wikimedia projects from other countries[18]. Additionally, we included the daily number of internet subscribers in India as a covariate, which is linearly interpolated from a quarterly series reported by Telecom Regulatory Authority of India.

4.2.2 Results. For both the online and TV campaigns for Hindi Wikipedia, we do not detect a significant change in page views. As we discuss below and in §5.1, this is likely a combination of factors: low impact and imprecise control series. The results from the BSTS models for the online campaign are in Figure 3 and TV campaign in Figure 4.

First we examine the results for the online video campaign that was targeted at the state of Madhya Pradesh. As before, we performed a 10-fold cross validation with the pre-intervention time

[10] http://www.internetlivestats.com/internet-users/india/

[11] https://economictimes.indiatimes.com/tech/internet/hindi-content-consumption-on-internet-growing-at-94-google/articleshow/48528347.cms

[12] https://meta.wikimedia.org/wiki/New_Readers/Raising_Awareness_in_India

[13] https://commons.wikimedia.org/wiki/File:Wikipedia_-_Ektara_(English_subtitles).webm

[14] TV data was collected by Eurodata TV via BARC in India.

[15] These Hindu festivals are picked because of their relative big impact on the treated time series. Their dates of each year are obtained from https://www.officeholidays.com/countries/india/index.php

[16] We selected the top 10 Wikimedia projects in India with the most page views, and Wikipedia of major Indian languages spoken by more than 4% of the population, according to 2011 census of India.

[17] https://www.trai.gov.in/release-publication/reports/performance-indicators-reports

[18] Countries that contribute more than 5% of Hindi Wikipedia page views, countries whose official language is Hindi, and other nearby countries.

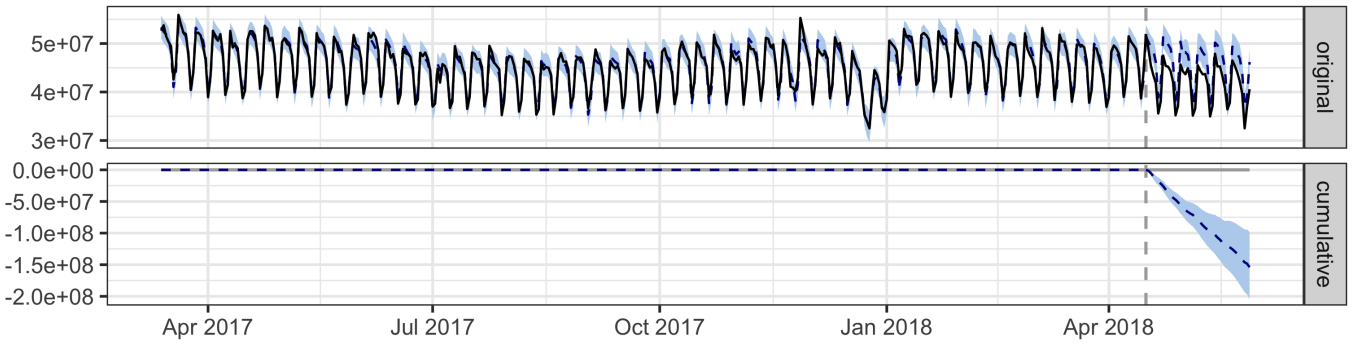


Figure 2: Impact of the page preview feature on 6 weeks of English Wikipedia internally-referred page views. Vertical dashed line represents the date of the roll-out (17 April 2018). Shaded areas indicate 95% credible intervals. The first panel shows the data (black solid line) and counterfactual prediction (blue dashed line) for the post-intervention period. The second panel sums the difference between observed data and counterfactual predictions—i.e. point-wise causal effect as estimated by the model, resulting in a plot of the cumulative effect of the intervention. The figures of point-wise causal effect are removed in this paper for space consideration.

Figure 3: Impact of the online campaign in 6 weeks on Hindi Wikipedia externally-referred page views from Madhya Pradesh. Vertical dashed line represents the start date of the campaign 3 April 2018.

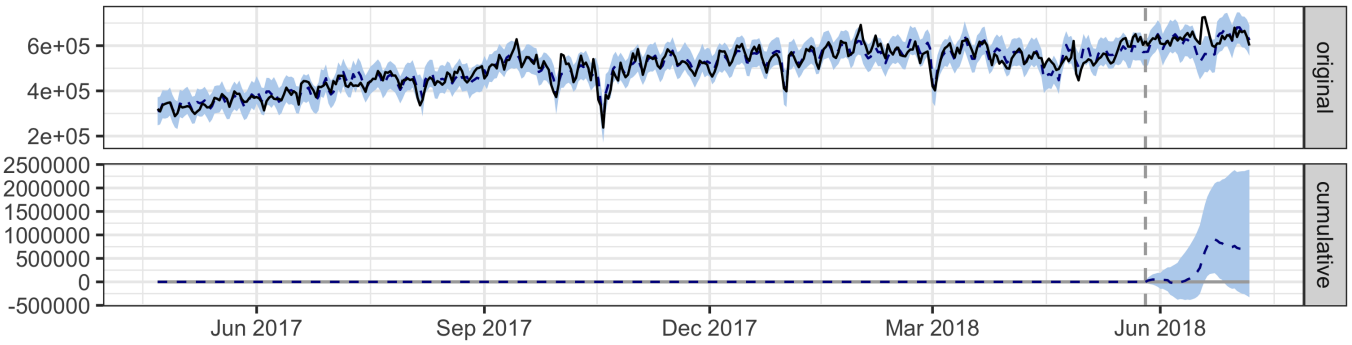


Figure 4: Impact of the TV campaign in 6 weeks on Hindi Wikipedia externally-referred page views from all of India. Vertical dashed line represents the date of the campaign 27 May 2018.

series and predict 6-week's daily page views in each fold. The average holdout MAPE is 7.6%.

As Figure 3 indicates, no significant impact on page views was detected following the intervention. While there does appear to be a downward trend in page views, zero change is still within the 95% credible interval. Page views to hi-wiki from the states of Rajasthan and Chhattisgarh, both of which border Madhya Pradesh, are the most predictive control series in the model, with average

standardized coefficients of 0.23 and 0.13 respectively. The posterior probabilities that their coefficients are different from zero are greater than 95%.

Next we turn to the results for the country-wide TV campaign. The average holdout MAPE from the cross validation is 10.2%. Figure 4 shows the results from the BSTS model. As the graphs indicate, no significant impact on page views was detected in the first 3 weeks following the intervention. There was a bump in the 4th week after the campaign, but it is most likely to be the result of an unknown event. Overall, we did not detect significant impact in 6 weeks. The number of internet subscribers in India, hi-wiki page views from the United States, Bengali Wikipedia page views from Bangladesh and English Wiktionary pageviews from Nepal are the most predictive control series in the model, with an average standardized coefficient of 0.84, 0.35, 0.14 and 0.16 respectively. The posterior probabilities that their coefficients are different from zero are greater than 95%.

We did not include the page views to other Wikipedia language editions that also were geolocated to India—e.g., page views to en-wiki from India—in the set of control series because it may violate the independence assumption. Most people in India are multilingual, so if our brand awareness was affected by the campaign, the impact would likely be revealed on page views of other Wikipedia language editions from the target region as well. After seeing the relatively high MAPE (10.2%) of the TV campaign model, we tried to include the page views to other Wikipedia language editions and other Wikimedia projects from India into the model to see if they help. The average holdout MAPE decreased to 8.5%, but that model also does not detect significant impact.

5 DISCUSSION

5.1 Choosing a Control Time Series

A predictive control series is one of the most important aspects of a BSTS model and also the part of the model that is often most difficult to choose. From the page preview roll-out analysis (§4.1), we see that the BSTS model, with a well-chosen control time series, can effectively estimate the impact of a given external event. In that case, externally-referred and internally-referred page views for the same project are highly correlated, but only internally-referred page views were believed to be impacted by the page preview roll-out. For both German and English Wikipedia, the estimated causal impact was slightly conservative—i.e. lower in magnitude than expected based on the A/B tests—but still quite close to the expected impact. We had less success selecting an effective control series for the Hindi video campaign. Furthermore, the fact that the online campaign model for Madhya Pradesh had less error than the TV campaign model for all of India raises questions about how factors like geography or language edition affect the predictive power of a control series for Wikipedia.

To better understand the power of different types of control series, we tested four additional control time series models for the Hindi analysis. All are trained on 400 days of daily externally-referred page views and evaluated via 10-fold cross validation on 6 weeks of page views prior to the intervention (3 April 2018). All models include day-of-week and monthly seasonality, holiday effects, local level trends, and a set of control time series but no further covariates. The control series for each model and its respective validation error are shown in Table 1.

We see that the control time series that are from the same language edition as the treated time series (Models 1 and 3) have a consistently lower error than the models that are from different language edition but the same geographic region (Models 2 and 4). This indicates that language edition plays a more important role than geographic region in the page view trends on Wikipedia. Comparing Model 1 and 3 where the control and treated series are from the same language edition, page views between states within the same country are more predictive of each other (adding these controls into the model decrease the MAPE from 11.54% to 7.54%) than page views between different countries (adding them into the model only decrease the MAPE from 7.92% to 7.22%), which indicates that while language appears to be most important, country borders are still a highly salient aspect of page view trends on Wikipedia. Further research would be needed to understand how these effects play out in other language communities and the inter-relatedness of different countries and language pairs.

6 FUTURE WORK AND LIMITATIONS

While this work has a number of limitations, as we lay out below, we believe it lays the groundwork for exploring more standardized methods of predicting trends such as page views on Wikipedia with the goal of understanding the effect of external events. Limitations for this work largely relate to temporal evolution of impact, data pre-processing, prior distributions of parameters in BSTS, and the need for additional experiments.

In this work, we focus only on the cumulative effect by the end of the post-intervention period—its existence and magnitude—without discussing the temporal evolution of an impact. In practice, how an effect evolves over time, especially its onset and decay structure, is often a valuable question as well. The point-wise effect from BSTS reflects the temporal evolution and future implementation should consider analyzing this result.

Small volume Wikipedia editions such as hi-wiki are more sensitive to undetected bot behavior, which can cause anomalies in page-view data. Anomalies in the prediction or post-intervention period would increase the error rate of validation, or the model might detect an impact that is unrelated to the known intervention. When the number of control series is very large, removing outliers manually is not feasible and thus requires a robust algorithm to detect and adjust outliers while preserving those known "outliers" such as holiday effects. It is possible that further pre-processing would also provide benefits—e.g., including more holidays, removing seasonal patterns in predictors before fitting the model, more extensive grid search for parameters like the length of pre-intervention period.

We were expecting that spike-and-slab prior in the BSTS would prevent over-fitting by forcing the coefficients of poor predictors to zero, so we would at least have predictions not worse than that of a model which only contains the historical information of the treated series itself. Contrary to this expectation, Model 4 from Table 1 (predicting hi-wiki page views in India) shows that including control series from other Wikipedia language editions and other Wikimedia projects within India actually added noise to the prediction. To solve

Table 1: Comparison of the predictive power of four sets of control time series for the Hindi Wikipedia campaign, of which the treated series and the control series either share the same geographic region or the same Wikipedia language edition.

	Treated Series (lang; region)	Control Series (lang; region)	Avg MAPE	Avg MAPE (No Control Series)
Model 1	hi-wiki page views; Madhya Pradesh	hi-wiki page views; other states of India	7.54%	11.54%
Model 2	hi-wiki page views; Madhya Pradesh	other wikis page views; Madhya Pradesh	9.31%	11.54%
Model 3	hi-wiki page views; all of India	hi-wiki page views; other countries	7.22%	7.92%
Model 4	hi-wiki page views; all of India	other wikis page views; all of India	9.14%	7.92%

this problem, we can further tune the hyper-parameter that controls the expected model size—the expected number of coefficients that are different from zero—so that when most of the predictors do not have enough predictive power, we can lower the expected model size and force more coefficients to be zero (we set the expected model size to be 10% of total number of controls in the model but not greater than 5 in this work).

Finally, future work should continue to explore these models in more contexts. This would hopefully provide guidance for how to select control time series—e.g., which pairs of regions and language editions (or even other Wikimedia projects) are predictive, what is the best way to split Wikipedias into control and treatment articles, how to take advantage of more granular information as with the internal/external referrer information. This would also hopefully provide guidance for how to set priors in BSTS models—e.g., a prior likelihood of relationships between treatment and control time series or prior standard deviation of the Gaussian random walk of the trend models (conservatively, we use a non-informative prior for the former and 0.01 for the latter). In future analyses, we can increase the prior inclusion probabilities for control series that are likely to be correlated with the treated series, and increase the prior standard deviation for the trend models if we believe the volatility of residuals would be large after regressing out known predictors.

ACKNOWLEDGMENTS

We would like to thank Tilman Bayer, Mikhail Popov, Leila Zia, Kate Zimmerman and Jon Katz for guidance and support as well as our anonymous reviewers for their feedback.

REFERENCES

[1] Alexa. 2018. *wikipedia.org Traffic Statistics.* Technical Report. Alexa. https://www.alexa.com/siteinfo/wikipedia.org

[2] Joshua D. Angrist and Pischke Jörn-Steffen. 2009. *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton University Press.

[3] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. 2002. How Much Should We Trust Differences-in-Differences Estimates? (2002). https://doi.org/10.3386/w8841

[4] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9, 1 (2015), 247–274. https://doi.org/10.1214/14-aoas788

[5] Jennifer L. Castle, Xiaochuan Qin, and W. Robert Reed. 2009. How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms, by Jennifer L. Castle; Xiaochuan Qin; W. Robert Reed. https://ideas.repec.org/p/cbt/econwp/09-13.html

[6] Justin Clark, Robert Faris, and Rebekah Heacock Jones. 2017. *Analyzing Accessibility of Wikipedia Projects Around the World.* Berkman Klein Center for Internet & Society Research Publication.

[7] Ed Erhart. 2019. Wikipedia's most-popular articles of 2018 show that pop culture rules over us all. https://wikimediafoundation.org/2019/01/02/wikipedias-most-popular-articles-of-2018-show-that-pop-culture-rules-over-us-all/

[8] E. I. George and R. E. McCulloch. 1997. Approaches for Bayesian variable selection. *Statistica Sinica* 7 (1997), 339–374.

[9] Satdeep Gill and Zack McCune. 2018. How we're building awareness of Wikipedia in India. https://blog.wikimedia.org/2018/04/03/building-awareness-wikipedia-india/

[10] Andrew C. Harvey. 1991. *Forecasting, structural time series models and the Kalman filter.* Cambridge Univ. Press.

[11] D. Madigan and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* 89 (1994), 1535–1546.

[12] Márton Mestyán, Taha Yasseri, and János Kertész. 2013. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one* 8, 8 (2013), e71226.

[13] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. 2013. Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports* 3 (2013), 1801.

[14] Daniel Cheng Moyer, Samuel L Carson, Thayne Keegan Dye, Richard T Carson, and David Goldbaum. 2015. Determining the Influence of Reddit Posts on Wikipedia Pageviews. In *Ninth International AAAI Conference on Web and Social Media.*

[15] Jonathon W Penney. 2016. Chilling effects: Online surveillance and Wikipedia use. *Berkeley Tech. LJ* 31 (2016), 117.

[16] Reid Priedhorsky, Dave Osthus, Ashlynn R Daughton, Kelly R Moran, Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, and Sara Y Del Valle. 2017. Measuring global disease with Wikipedia: Success, failure, and a research agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* ACM, 1812–1834.

[17] Steven L Scott. 2017. Fitting Bayesian structural time series with the bsts R package. http://www.unofficialgoogledatascience.com/2017/07/fitting-bayesian-structural-time-series.html

[18] Steven L Scott and Hal R Varian. 2013. *Bayesian Variable Selection for Nowcasting Economic Time Series.* Working Paper 19567. National Bureau of Economic Research. https://doi.org/10.3386/w19567

[19] Steven L Scott and Hal R Varian. 2013. Predicting the present with bayesian structural time series. *Available at SSRN 2304426* (2013).

[20] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 1591–1600.

[21] Marijn ten Thij, Yana Volkovich, David Laniado, and Andreas Kaltenbrunner. 2012. Modeling and predicting page-view dynamics on Wikipedia. *CoRR abs/1212.5943* (2012).

[22] Giorgino Toni. 2009. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software* 31 (2009). https://doi.org/10.18637/jss.v031.i07

[23] Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 566.

[24] Xiaoquan Michael Zhang and Feng Zhu. 2011. Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101, 4 (2011), 1601–15.

The Third Workshop on Women in Web Data Science (WinDS'19) Chairs' Welcome & Organization Listing

It is our great pleasure to welcome you to the *WWW 2019 Third Women in Web Data Science (WinDS'19) Workshop*. This half-day workshop brings together female faculties, graduate students, research scientists, and industry researchers for an opportunity to connect, exchange ideas, and learn from each other in the field of Data Science. Underrepresented minorities, graduates, and undergraduates interested in pursuing data science and machine learning research are encouraged to participate. While most presenters are women, everybody is invited to attend.

In order to build a community of Women in Web Data Science at *WWW 2019* to share knowledge, experiences, and drawbacks in the field, the workshop features keynote talks, accepted long paper presentations, and lighting talks from female first authors. Each submitted paper was peer-reviewed by the workshop program committee. The workshop is organized into two sessions, each including a keynote talk, accepted paper presentations and lighting talks. We are going to have two keynote talks, one from academia and one from industry:

1. *Anna Lisa Gentile (IBM Research Almaden):* Information Extraction with Humans in the Loop.

2. *Lise Getoor (University of California – Santa Cruz):* Responsible Data Science.

We believe that the program provides a good balance between several trending topics in Web Data Science such as social media analysis, online misinformation, misbehavior in social media, natural language processing, open data analysis, information retrieval, healthcare, recommender systems, etc. Hence we hope that you will find the workshop program interesting, providing you with a valuable opportunity to learn and share ideas and experiences with other researchers and practitioners from institutions around the world.

Ana Paula Appel
Workshop Chair
IBM Research

Marisa Vasconcelos
Workshop Chair
IBM Research

Francesca Spezzano
Workshop Chair
Boise State University

Célia Talma Gonçalves
Workshop Chair
CEOS.PP/ISCAP/P.PORTO & LIACC

THE WEB CONFERENCE

Third Workshop on Women in Web Data Science (WinDS'19) Organization

Workshop Chairs: Ana Paula Appel *(IBM Research, Brazil)*
Marisa Vasconcelos *(IBM Research, Brazil)*
Francesca Spezzano *(Boise State University, USA)*
Célia Talma Gonçalves *(CEOS.PP/ISCAP/P.PORTO & LIACC, Portugal)*

Program Committee: Andrea Britto Mattos Lima *(IBM Research, Brazil)*
Carmela Comito *(National Research Council, Italy)*
Homa Hosseinmardi *(University of Southern California, USA)*
Ana Carolina Lorena (*Federal University of São Paulo, Brazil)*
Maysa M G Macedo *(IBM Research, Brazil)*
Élia Yathie Matsumoto *(University of São Paulo, Brazil)*
Mirjana Mazuran *(INRIA Saclay, France)*
Ida Mele *(University of Italian Switzerland, Switzerland)*
Maria Camila Nardini Barioni *(Federal University of Uberlandia, Brazil)*
Alexandra Olteanu *(Microsoft Research, Canada)*
Efi Papatheocharous *(RISE Research Institutes of Sweden, Sweden)*
Arti Ramesh *(Binghamton University, USA)*

Information Extraction with Humans in the Loop

Anna Lisa Gentile
annalisa.gentile@ibm.com
IBM Research Almaden, CA, US

ABSTRACT

Information Extraction (IE) techniques enables us to distill Knowledge from the abundantly available unstructured content. Some of the basic IE methods include the automatic extraction of relevant entities from text (e.g. places, dates, people, ...), understanding relations among them, building semantic resources (dictionaries, ontologies) to inform the extraction tasks, connecting extraction results to standard classification resources. IE techniques cannot decouple from human input - at bare minimum some of the data needs to be manually annotated by a human so that automatic methods can learn patterns to recognize certain type of information. The human-in-the-loop paradigm applied to IE techniques focuses on how to better take advantage of human annotations (the recorded observations), how much interaction with the human is needed for each specific extraction task.

ACM Reference Format:
Anna Lisa Gentile. 2019. Information Extraction with Humans in the Loop. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316478

TALK OUTLINE

In this talk I will describe various experiments of the human-in-the-loop model on various IE tasks, such as (i) building dictionaries from text corpora in various languages [1]; (ii) extracting mentions of adverse drug reaction from text and matching them to a reference ontology [2]; (iii) relation extractions, e.g. automatically identifying from the text which drug is causing which adverse drug reaction [3].

SPEAKER'S BIO

Dr Anna Lisa Gentile (https://w3id.org/people/annalisa) is a Research Staff Member at IBM Research Almaden. Her research is principally focused on studying methods and techniques for semantic annotating unstructured and semi-structured content. Her main Research Areas are Information Extraction (IE), Natural Language Processing (NLP) and Semantic Web. She obtained her PhD with a thesis on Named Entity Disambiguation at the University of Bari, Italy in 2010. She has published more than 50 peer-reviewed scientific publications including papers at major venues such as LREC, EMNLP, ESWC and ISWC. She has been serving as Organizing Committee member for conferences such as ISWC, ESWC, WWW amongst many others and organized workshop series such as *LD4IE*

on Linked Data for Information Extraction (http://w3id.org/ld4ie) and *HumBL* on Augmenting Intelligence with Bias-Aware Humans-in-the-Loop (http://w3id.org/huml).

REFERENCES

[1] Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. 2017. Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop. In *K-CAP 2017*, Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo (Eds.). ACM, 24:1–24:8. https://doi.org/10.1145/3148011.3148021

[2] Kenneth Clarkson, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, Joseph Terdiman, and Steve Welch. 2018. User-Centric Ontology Population. In *ESWC 2018 (Lecture Notes in Computer Science)*, Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam (Eds.), Vol. 10843. Springer, 112–127. https://doi.org/10.1007/978-3-319-93417-4_8

[3] Ismini Lourentzou, Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, and Steve Welch. 2018. Mining Relations from Unstructured Content. In *PAKDD 2018*. 363–375. https://doi.org/10.1007/978-3-319-93037-4_29

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316478

Responsible Data Science

Keynote

Lise Getoor
University of California, Santa Cruz
getoor@soe.ucsc.edu

ABSTRACT

Data science is an emerging discipline that offers both promise and peril. *Responsible data science* refers to efforts that address both the technical and societal issues in emerging data-driven technologies. How can machine learning and AI systems reason effectively about complex dependencies and uncertainty? Furthermore, how do we understand the ethical and societal issues involved in data-driven decision-making? There is a pressing need to integrate algorithmic and statistical principles, social science theories, and basic humanist concepts so that we can think critically and constructively about the socio-technical systems we are building. In this talk, I will overview this emerging area.

ACM Reference Format:
Lise Getoor. 2019. Responsible Data Science: Keynote. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316479

Lise Getoor is a professor in the Computer Science Department at UC Santa Cruz and director of the Data, Discovery and Decisions (D^3) Data Science Research Center at the University of California, Santa Cruz. Her research areas include machine learning, data integration and reasoning under uncertainty, with an emphasis on graph and network data. She has over 200 publications, including 12 best paper awards, and over 19,000 citations. She is a Fellow of the Association for Artificial Intelligence, has served as an elected board member of the International Machine Learning Society, and the Computing Research Association (CRA). This past fall she was a visiting researcher in the Fairness, Accountability, Transparency and Ethics research group at Microsoft Research, NYC. She received her PhD from Stanford University in 2001, her MS from UC Berkeley, and her BS from UC Santa Barbara, and was a professor at the University of Maryland, College Park from 2001-2013.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316479

Lightning Talk - Looking for the Movie Seven or Sven from the Movie Frozen?

A multi-perspective strategy for recommending queries for children

Oghenemaro Anuyah
People and Information Research Team (PIReT)
Department of Computer Science — Boise State University — Boise, Idaho
oghenemaroanuyah@u.boisestate.edu

Ion Madrazo Azpiazu
People and Information Research Team (PIReT)
Department of Computer Science — Boise State University — Boise, Idaho
ionmadrazo@u.boisestate.edu

Nevena Dragovic
People and Information Research Team (PIReT)
Department of Computer Science — Boise State University — Boise, Idaho
nevenadragovic@u.boisestate.edu

Maria Soledad Pera
People and Information Research Team (PIReT)
Department of Computer Science — Boise State University — Boise, Idaho
solepera@boisestate.edu

ABSTRACT

As one of the largest communities that search for online resources, children are introduced to the Web at increasingly young ages [1]. However, popular search tools are not explicitly designed with children in mind nor do their retrieved results explicitly target children. Consequently, many young users struggle in completing successful searches, especially since most search engines (**SE**) do not directly support, or offer weak support, for children's inquiry approaches [2]. Even though children, as inexperienced users, struggle with describing their information needs in a concise query, they still expect SEs to retrieve relevant information in response to their requirements. As part of their capabilities, SEs often suggest queries to aid users in better defining their information needs. In fact, a recent study conducted by Gossen et al. [3] shows that children pay more attention to suggested queries than adults. Unfortunately, these suggestions are not specifically tailored towards children and thus need improvement [5]. While there exist multiple query suggestion modules, only few specifically target children. To address this problem, along with a need for more children-related tools, we rely on *ReQuIK* (**Re**commendations based on **Qu**ery **I**ntention for **K**ids), a query suggestion module tailored towards 6-to-13 year old children (introduced in [4]). *ReQuIK* informs its suggestion process by applying (i) a strategy based on *search intent* to capture the purpose of a query [1], (ii) a ranking strategy based on a *wide* and *deep neural network* that considers both raw text and traits commonly associated with kid-related queries, (iii) a filtering strategy based on the *readability* levels of documents potentially retrieved by a query to favor suggestions that trigger the retrieval of documents matching children's reading skills, and (iv) a content-similarity strategy to ensure *diversity* among suggestions.

For assessing the quality of the system, we conducted initial offline and online experiments based on 591 queries written by 97 children, ages 6 to 13. The results of this assessment verified the correctness of *ReQuIK*'s recommendation strategy, the fact that it provides suggestions that appeal to children and *ReQuIK*'s ability to recommend queries that lead to the retrieval of materials with readability levels that correlate with children's reading skills

To the best of our knowledge, *ReQuIK* is the only available system that can be coupled with SEs to generate query recommendations for children, favoring those that lead to easier-to-read, child-related resources, which can improve SEs' performance. The design of the proposed tool explicitly considers different patterns children use while searching the Web to adequately capture the intended meaning of their original queries. For example, if a child submits the query *"elsa"*, *ReQuIK* aims to prioritize query suggestions such as *"elsa coloring papers"* or *"elsa dress up games"* that correlate better with topics of interest to children rather than *"elsa pataky"*, as suggested by Google, which is more appealing to mature users. Other contributions of our work include a novel ranking model inspired by a deep-and-wide architecture that, while successfully applied for ranking purposes [6], has never been used in the query suggestions domain; a strategy to overcome the lack of queries written by children by taking advantage of general purpose children-oriented phrases; and a newly-created dataset [4].

CCS CONCEPTS

• **Information systems → Query suggestion**; • **Social and professional topics → Children**.

KEYWORDS

Query suggestions; children; search intent; dataset

ACM Reference Format:
Oghenemaro Anuyah, Ion Madrazo Azpiazu, Nevena Dragovic, and Maria Soledad Pera. 2019. Lightning Talk - Looking for the Movie Seven or Sven from the Movie Frozen?: A multi-perspective strategy for recommending queries for children. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316475

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316475

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation Award Number 1565937.

REFERENCES

[1] Nevena Dragovic, Ion Madrazo Azpiazu, and Maria Soledad Pera. 2016. Is Sven Seven?: A Search Intent Module for Children. In *ACM SIGIR*. 885–888.

[2] Elizabeth Foss, Allison Druin, Robin Brewer, Phillip Lo, Luis Sanchez, Evan Golub, and Hilary Hutchinson. 2012. Children's search roles at home: Implications for designers, researchers, educators, and parents. *JASIST* 63, 3 (2012), 558–573.

[3] Tatiana Gossen, Juliane Höbel, and Andreas Nürnberger. 2014. A comparative study about children's and adults' perception of targeted web search engines. In *SIGCHI*. 1821–1824.

[4] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. 2018. Looking for the Movie Seven or Sven from the Movie Frozen?: A Multi-perspective Strategy for Recommending Queries for Children. In *ACM CHIIR*. ACM, 92–101.

[5] Arif Usta, Ismail Sengor Altingovde, Ibrahim Bahattin Vidinli, Rifat Ozcan, and Özgür Ulusoy. 2014. How k-12 students search for learning?: analysis of an educational search engine log. In *ACM SIGIR*. 1151–1154.

[6] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational Context for Ranking in Personal Search. In *WWW*. 1531–1540.

Lightning Talk - Humor Recognition in Russian Language

Valeriia Baranova-Bolotova
Ural Federal University
Yekaterinburg, Russian Federation
lurunchik@gmail.com

Vladislav Blinov
Ural Federal University
Yekaterinburg, Russian Federation
vladislav.blinov@urfu.ru

Pavel Braslavski
Ural Federal University
Yekaterinburg, Russian Federation
pbras@yandex.ru

ABSTRACT

In this lighting talk paper, we present a dataset of jokes in Russian and deep learning model for solving humor recognition task. The new large dataset was collected from various online resources and complemented carefully with unfunny texts with similar lexical properties. In total, there are more than 300,000 short texts, which is significantly larger than any previous humor-related corpus. Manual annotation of 2,000 items proved the reliability of corpus construction approach. Further, we applied language model fine-tuning for text classification and obtained an F1 score of 0.91, which constitutes a considerable gain over baseline methods.

KEYWORDS

natural language processing; humor recognition; neural networks; language modelling

ACM Reference Format:
Valeriia Baranova-Bolotova, Vladislav Blinov, and Pavel Braslavski. 2019. Lightning Talk - Humor Recognition in Russian Language. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316473

1 DATA

Our goal was to expand Stierlitz dataset of Russian jokes with jokes and non-jokes lexically similar to them.

Firstly, we collected more than 1M jokes from multiple humorous public pages from the social network *VK*[1] and from the website *anekdot.ru*[2]. Further, we downloaded 10M posts from a large online forum *E1.ru*[3], and indexed them with Elastic[4] and returned the first post with Jaccard similarity less than 0.4 to the collection from the BM25-ranked list of matching forum posts for each existing joke. In total, we had 314,269 jokes with their non-jokes pairs. We also used Puns [1] dataset of 213 puns for evaluation.

To verify our automatically created collection, we conducted an evaluation using an online interface, where 2,000 random jokes and non-jokes were assessed on a 3-point scale: 'not a joke', 'an unfunny joke' and 'a joke'. More than 100 volunteers took part in evaluation and 1,877 examples were labeled by at least three assessors. Majority voting resulted in 94% of non-jokes marked as 'not a joke' and 95% of jokes marked as either 'an unfunny joke' or 'a joke', which demonstrates a good performance of the automatic procedure.[5]

[1]http://vk.com/ - the largest Russian social network
[2]https://www.anekdot.ru/ - the oldest website of jokes on the Russian Web
[3]https://www.e1.ru/talk/forum/
[4]https://www.elastic.co/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316473

Table 1: Detection quality – F1 scores on Stierlitz, Fun, Gold, and recall on jokes-only Puns dataset

Model	Stierlitz Test	Fun Test	Gold	Puns
Trained on Stierlitz train				
Baseline SVM	0.91	0.71	0.64	0.73
Stierlitz SVM [1]	0.88	0.74	0.64	0.7
ULMFun	**0.97**	0.77	0.66	**0.92**
Trained on Fun train				
Baseline SVM	0.8	0.8	0.8	0.43
ULMFun	0.92	**0.91**	0.88	**0.92**

2 CLASSIFICATION METHODS

Initially, we pretrained a language model on 10M online forum texts for 15 epochs with architecture and parameters directly transferred from [2]. Texts were tokenized using unigram subword tokenization method implemented in SentencePiece library [3] with the vocabulary size of 100,000. Further, we used either Stierlitz or Fun dataset to fine-tune the model for five epochs. Finally, we replaced the target task with humor classifier by augmenting the model, further referred to as ULMFun, with linear blocks and trained the model with gradual unfreezing followed by 14 consecutive epochs. We chose linear SVM classifier on top of *tf.idf* features as a baseline.

3 RESULTS

The goals of the experiment were to estimate the impact of the increased dataset size and its construction methods; to introduce a strong baseline based on deep neural network approach, to compare it with a baseline and published work, as well as to evaluate generalization abilities of the model. In the first series of experiments we trained a simple SVM baseline on *tf.idf* features and ULMFun model on Stierlitz train set. We tested the obtained models on held-out test sets of Stierlitz and Fun, as well as on smaller manually annotated Gold and Puns collections. In addition, we were able to apply the best model from [1] to the test data. In the second series, we trained simple baseline and ULMFun on a larger Fun training set and evaluated these two models on the same test data as in the previous stage. Table 1 summarizes the results. ULMFun significantly outperforms Stierlitz SVM and our baseline. The lower part of Table 1 shows results of baseline model and ULMFun

[5]For example, manual verification of one-liners dataset [4] revealed 9% of noise.

trained on a larger Fun training set. As expected, more data significantly improves classification quality on Fun test set in case of both methods.

REFERENCES

[1] Anton Ermilov, Natasha Murashkina, Valeria Goryacheva, and Pavel Braslavski. 2018. Stierlitz Meets SVM: Humor Detection in Russian. In *Artificial Intelligence and Natural Language*. 178–184.

[2] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 328–339.

[3] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 66–71.

[4] Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 531–538.

Signed Link Prediction with Sparse Data: The Role of Personality Information

Ghazaleh Beigi
Computer Science and Engineering
Arizona State University
gbeigi@asu.edu

Suhas Ranganath
Walmart Labs
suhas.ranganath@walmartlabs.com

Huan Liu
Computer Science and Engineering
Arizona State University
huan.liu@asu.edu

ABSTRACT

Predicting signed links in social networks often faces the problem of signed link data sparsity, i.e., only a small percentage of signed links are given. The problem is exacerbated when the number of negative links is much smaller than that of positive links. Boosting signed link prediction necessitates additional information to compensate for data sparsity. According to psychology theories, one rich source of such information is user's personality such as optimism and pessimism that can help determine her propensity in establishing positive and negative links. In this study, we investigate how personality information can be obtained, and if personality information can help alleviate the data sparsity problem for signed link prediction. We propose a novel signed link prediction model that enables empirical exploration of user personality via social media data. We evaluate our proposed model on two datasets of real-world signed link networks. The results demonstrate the complementary role of personality information in the signed link prediction problem. Experimental results also indicate the effectiveness of different levels of personality information for signed link data sparsity problem.

CCS CONCEPTS

• **Information systems → Social recommendation**; • **Human-centered computing → Social engineering (social sciences)**.

KEYWORDS

Signed Link Prediction; Data Sparsity; Personality Information; Optimism; Pessimism

ACM Reference Format:
Ghazaleh Beigi, Suhas Ranganath, and Huan Liu. 2019. Signed Link Prediction with Sparse Data: The Role of Personality Information. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316469

1 INTRODUCTION

The rapid growth of social media has led to an increased amount of user generated data and accordingly the impediment of finding reliable information [1, 7]. Positive links (e.g., trust and friendship relations) play an important role in helping online users find relevant and credible information [37]. They have also been demonstrated to benefit many social media applications including recommendation and information filtering [8, 41]. Similarly, negative links could help decision makers reduce uncertainty and vulnerability associated with decision consequences [15, 20, 26, 29]. Therefore, signed link prediction without negative links may result in a biased estimate of the effect of positive links [39]. Thus it is sensible to investigate both positive and negative links together in signed link prediction.

The problem of signed link prediction aims at inferring new positive/negative relations by leveraging existing ones. In recent years, the majority of the existing algorithms [14, 27] use the topological structures and the properties of the existing signed links to make predictions. However, the available explicit positive links are often sparse and follow power-law distributions [37]. The signed data sparsity problem gets worse as social media users tend to reveal more their positive disposition than their negative one; thus, negative links are often much sparser than positive links in a signed social network. To make a better signed link predictor, we need to overcome the data sparsity problem of the signed links.

As suggested by psychologists [4, 12], user's optimism and pessimism are important factors that determine her propensity in establishing the positive/negative social relations. According to Scheier et. al. [31], a person is defined as *optimist* when she is more likely to reinterpret negative events in a positive way and find meaning and growth in stressful situations. On the other hand, an individual is referred to as *pessimist* when she is pre-occupied only with the negative aspects of the environment and overlooks the positive aspects [31]. For example, optimistic users have better social functioning and relations. Therefore, they actively pursue social relationships and have higher chances in establishing positive links resulting in longer lasting friendships [17, 31]. In contrast, pessimists likely practice in the opposite way, i.e., having negative attitudes and expecting the worst of people and situations. Consequently, they often establish negative links with others [17, 33]. People's personal traits can be observed on social media and serve as a good indicator of their personality [18]. This is because (1) social media websites allow for free interaction and open exchange of viewpoints, and (2) social media data can be aggregated to establish normative behavior of individuals. Previous research [10] has shown the correlation between users' personality information and positive/negative links in social networks. In particular it shows that:

- Users with high optimistic behavior are more likely to establish and receive positive links than those with low optimism.
- Users who are more pessimistic are more likely to establish and receive negative links than those with low level of pessimism.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316469

These findings are in line with psychology research [17, 31, 33] and also suggest that user's personality, i.e., optimism and pessimism, may have potentials to alleviate the signed link data sparsity problem and improve the performance of signed link prediction. We use optimism and pessimism as a representative aspect of personality.

Although previous research [10] has taken the first steps to study the correlation between users' personality and positive/negative links in signed social networks, it is still unclear how such information could be modeled mathematically and incorporated for predicting the positive and negative links, whether the information could help solving signed link data sparsity problem and what the impact of personality information is on signed link prediction problem. In this paper, we study the problem of signed link prediction by exploiting personality information, in particular, users' optimism and pessimism. In particular, we investigate how to leverage such information for positive and negative link prediction, and then we propose a novel signed link prediction framework *SLP*. Our main contributions are as follows:

(1) Provide a principled way to model optimism and pessimism information mathematically;
(2) Propose the framework *SLP* which deploys personality information for the signed link prediction problem; and
(3) Evaluate extensively *SLP* on two datasets of real-world signed link networks and explore the impact of personality information on signed link prediction.

2 PROBLEM STATEMENT

We shall first assume $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ is the set of n users. We use the matrix $\mathbf{G} \in \mathbb{R}^{\mathbf{n}\times\mathbf{n}}$ to denote user-user positive and negative links where $\mathbf{G_{ij}} = \mathbf{1}$, $\mathbf{G_{ij}} = \mathbf{-1}$ and $\mathbf{G_{ij}} = \mathbf{0}$ represent a positive link, a negative link, and missing (a.k.a unknown) information from u_i to u_j respectively. We also measure users' optimism and pessimism personality information following previous research [10]–this is discussed in more details in the next section. We follow psychology literature and consider two *separate* measures for each user's optimism and pessimism as they should be conceptualized as two independent dimensions [10, 13]. This means that being optimistic does not necessarily mean not pessimistic. Let us now assume vectors $\mathbf{o} \in \mathbb{R}^{n\times 1}$ and $\mathbf{p} \in \mathbb{R}^{n\times 1}$ denote the users' optimism and pessimism respectively, where $\mathbf{o}_i \in [0, 1]$ is the optimism strength of u_i and $\mathbf{p}_i \in [0, 1]$ is her pessimism strength. The higher optimism (pessimism) score is, the more optimistic (pessimistic) u_i is.

The problem of signed link prediction by exploiting user's optimism and pessimism is then formally defined as follows:
Given users' optimism and pessimism $\mathbf{o}$ *and* $\mathbf{p}$*, and user-user existing positive and negative links matrix* $\mathbf{G}$*, we seek to learn a predictor* f *to find the unknown positive/negative link information by inferring new user-user positive/negative link matrix* $\widetilde{\mathbf{G}}$ *as follows:*

$$f : \{\mathbf{G}, \mathbf{o}, \mathbf{p}\} \rightarrow \{\widetilde{\mathbf{G}}\} \tag{1}$$

The problem of signed link prediction is different than its existing variants, positive link prediction [37] and sign prediction [42] and more challenging compared to them. In particular, we aim to predict the existence of a link between a pair of nodes and its sign. We also predict both positive and negative links simultaneously. Besides, the vast majority of the existing work for the problem of signed link prediction [27] leverage only the existing links between users. In our work, we seek to leverage additional resources such as user's personality to alleviate the data sparsity problem.

3 COMPUTATION OF PERSONALITY INFORMATION

Personality information is not readily available on social media. Individuals usually do not label themselves as optimistic or pessimistic. A conventional way of obtaining personality information is to directly ask people whether they expect outcomes in their lives to be good or bad, which is often seen to use psychological surveys designed for measuring an individual's optimism and pessimism (e.g., [32]). However, since social media data is large-scale, and mainly observational, it is impractical to ask each and every user for their personality information. The onus is, therefore, on us to find a sensible way to infer if a user is optimistic or pessimistic or neither. An indirect approach is to measure optimism and pessimism based on the idea that people's expectancies for the future stem from their interpretations of the past. Thus, past experiences can reflect an individual's levels of optimism or pessimism. With social media data, the question is how to define a computational measure of optimism or pessimism. To summarize, individuals do not explicitly offer their personality information, it is infeasible to ask a large number of users about such information, but individual social media users do leave their traces online. We ask if we can aggregate individual user's data and automatically figure out if a user is optimistic or not.

Scheier et. al. [31] defines optimism as re-interpreting negative events in a positive way and pessimism as preoccupying with the negative aspects and overlooking positive events. Following the psychology literature, user's feedback could be also used to estimate her optimism and pessimism as they are counterparts of each other [22, 23]. It is shown in [22, 23] that on social media websites, optimists are more willing to give more positive feedback while pessimists are more biased toward giving more negative feedback than usual. In previous research [10], this observation has been utilized to calculate users' optimism and pessimism by leveraging their feedback to different entities in social media (e.g. posts and items). In this work, we use two different scenarios to illustrate how personality scores can be computed. The two scenarios differ in their user feedback: (1) ratings given by users to items, or (2) opinions expressed by users towards each other.

3.1 Scenario 1: Ratings as Feedback

Here, feedbacks are users' ratings given to the items. Let $\mathcal{I}$ be the set of M items and assume r_{ik} denotes the rating score given from u_i to I_k with $r_{ik} = 0$ indicating that u_i has not rated I_k yet. Also, consider $\bar{r}_k$ as the average rating score of the k-th item rated by users. Following [10], we treat ratings less than some predefined threshold r_{th} as low and above it as high. We use $O_L(i)$ to denote the set of items with low average rating scores rated by u_i:

$$O_L(i) = \{I_k \mid r_{ik} \neq 0 \wedge \bar{r}_k \leq r_{th}\}$$

$O_{HL}(i)$ is further used to denote the set of items which have received high ratings from u_i, and meanwhile have low average scores.

$O_{HL}(i)$ can be formally defined as:

$$O_{HL}(i) = \{I_k \mid I_k \in O_L(i) \wedge r_{ik} > r_{th}\}$$

Intuitively, the more frequently user u_i has rated above the average, the more optimistic she is. Therefore, optimism score for u_i is defined as $\mathbf{o}_i = \frac{|O_{HL}(i)|}{|O_L(i)|}$ where $|.|$ is the size of the set [10].

Similarly, we use $\mathcal{P}_H(i)$ to denote the set of items with high average rating scores and rated by u_i,

$$\mathcal{P}_H(i) = \{I_k \mid r_{ik} \neq 0 \wedge \bar{r}_k > r_{th}\}$$

Let $\mathcal{P}_{LH}(i)$ denotes the subset of items from $\mathcal{P}_H(i)$, which are given low rates by u_i:

$$\mathcal{P}_{LH}(i) = \{I_k \mid I_k \in \mathcal{P}_H(i) \wedge r_{ik} \leq r_{th}\}$$

The pessimism score for u_i is defined as: $\mathbf{p}_i = \frac{|\mathcal{P}_{LH}(i)|}{|\mathcal{P}_H(i)|}$ [10].

3.2 Scenario 2: Opinions as Feedback

Here, feedbacks are users' opinions they expressed towards each other and individuals' personality is defined based on their positive/negative opinions [10]. Following the work of [10], we create user-user positive and negative opinion matrices $\mathbf{P}$ and $\mathbf{N}$ by computing the number of positive or negative opinions users express toward each other. Let $\overline{P}$ and $\overline{N}$ be the average of positive and negative opinions between all pairs of users, respectively. We also define $\overline{\mathbf{P}}_j$ and $\overline{\mathbf{N}}_j$ as the average of positive and negative opinions received by u_j. Further, we define $O_L(i)$, as a set of users u_j who have received positive emotions from u_i, but at the same time, have received more negative emotions than the average of negative opinions in the network, i.e. they are among the worst people in the network:

$$O_L(i) = \{u_j \mid \mathbf{P}_{ij} \neq 0 \wedge \overline{\mathbf{N}}_j > \overline{N}\}$$

$O_{HL}(i)$ denotes the set of users u_k who belong to $O_L(i)$ and have received more positive emotions from u_i than $\overline{\mathbf{P}}_k$,

$$O_{HL}(i) = \{u_k \mid u_k \in O_L(i) \wedge \mathbf{P}_{ik} > \overline{\mathbf{P}}_k\}$$

Intuitively, the more frequently u_i has given positive emotions to the worst users in the network, the more optimistic she is [10]. Thus, the optimism score for u_i is defined as as $\mathbf{o}_i = \frac{|O_{HL}(i)|}{|O_L(i)|}$.

Likewise, we define $\mathcal{P}_H(i)$, as a set of users u_j who have received negative emotions from u_i, but at the same time, have received more positive emotions than the average in the network, i.e. they are better than the average,

$$\mathcal{P}_H(i) = \{u_j \mid \mathbf{N}_{ij} \neq 0 \wedge \overline{\mathbf{P}}_j > \overline{P}\}$$

We define $\mathcal{P}_{LH}(i)$ to denote the set of users u_k who belong to $\mathcal{P}_H(i)$ and have received more negative emotions from u_i than $\overline{\mathbf{N}}_k$,

$$\mathcal{P}_{LH}(i) = \{u_k \mid u_k \in \mathcal{P}_H(i) \wedge \mathbf{N}_{ik} > \overline{N}_k\}$$

Pessimism of u_i could be similarly defined as: $\mathbf{p}_i = \frac{|\mathcal{P}_{LH}(i)|}{|\mathcal{P}_H(i)|}$ [10].

4 THE PROPOSED FRAMEWORK - SLP

Readily available information such as user's personality and its impact on the formation of signed links, motivates us to exploit it to overcome the inherent sparsity of the signed links. We model this information mathematically and then incorporate it for predicting positive and negative links. Here, we introduce our approach for modeling user's optimism and pessimism and then detail the proposed method *SLP* for signed link prediction.

4.1 Basic Model for Signed Link Prediction

Users usually establish signed links with only a few set of other users. This results in very sparse and low-rank networks. Therefore, low-rank approximation methods could be deployed for modeling signed links [21]. Moreover, the work of [21] showed that weak structural balance in signed networks leads to a low-rank approximation method for modeling the network. Let $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_n]^\top \in \mathbb{R}^{n \times d}$ be the low-rank latent representations of users in $\mathcal{U}$ where $\mathbf{U}_i \in \mathbb{R}^d$, $d \ll n$ is the low-rank latent vector representation of u_i. The matrix factorization model seeks a low-rank representation of $\mathbf{U}$ via solving the following optimization problem:

$$\min_{\mathbf{U},\mathbf{V}} \ \|\mathbf{W} \odot (\mathbf{G} - \mathbf{U}\mathbf{V}\mathbf{U}^\top)\|_{\mathbf{F}}^2 + \lambda_1 ||\mathbf{U}||_{\mathbf{F}}^2 + \lambda_2 ||\mathbf{V}||_{\mathbf{F}}^2, \qquad (2)$$

where $||\ .\ ||_F$ is the Frobenius norm of a matrix, $\odot$ is Hadamard product, and $\mathbf{V} \in \mathbb{R}^{d \times d}$ captures the correlations among user latent representations. $\mathbf{W}_{ij}$ controls the contribution of $\mathbf{G}_{ij}$ in the learning process. A typical choice of $\mathbf{W}$ is to set $\mathbf{W}_{ij} = 1$ if $\mathbf{G}_{ij} \neq 0$, and $\mathbf{W}_{ij} = 0$ otherwise. Two smoothness regularization terms are also added to avoid over-fitting where λ_1 and λ_2 are non-negative regularization parameters on $\mathbf{U}$ and $\mathbf{V}$, respectively. This standard model is very flexible to add prior knowledge from side information. In the following subsection, we introduce how to incorporate user personality information into the standard model.

4.2 Modeling User Personality Information

The analysis in the previous research [10] introduces two important findings according to users' optimism and pessimism: (a) users who are more optimistic tend to create and receive more positive links in comparison to the users with low level of optimism; and (b) users with more pessimism personality are more likely to create and receive more negative links compared to the users who are less pessimistic. To model the user's optimism and pessimism effect, we consider the following two cases for each pair of users $\langle u_i, u_j \rangle$:

- *Case 1:* $\mathbf{o}_i - \mathbf{o}_j > t_o$ and $d_i > d_j + \gamma_{ij}$;
- *Case 2:* $\mathbf{p}_j - \mathbf{p}_i > t_p$ and $d_j + \delta_{ij} < d_i$;

where t_o and t_p are thresholds to consider the significant optimism and pessimism difference between u_i and u_j, respectively. We use γ_{ij} and δ_{ij} to make the expected total degree comparisons flexible for u_i and u_j.

In *Case 1*, u_i has a higher level of optimism and u_j has a lower level of optimism, then $\mathbf{o}_i - \mathbf{o}_j > t_o$, which suggests that u_i is more likely to receive positive links in comparison to u_j. This results in a higher degree for u_i, i.e., $d_i > d_j + \gamma_{ij}$. To account for this, we impose a penalty if $d_j - d_i + \gamma_{ij}$ is lesser than 0, when $\mathbf{o}_i - \mathbf{o}_j > t_o$. Similarly, in *Case 2*, if u_j has a higher pessimistic level while u_i has a lower pessimism, then $\mathbf{p_j} - \mathbf{p_i} > \mathbf{t_p}$. This suggests that u_j is more

likely to receive negative links in comparison to u_i which results in a higher negative degree (and consequently a lower total degree) for u_j, i.e., $d_j + \delta_{ij} < d_i$. Therefore, we impose a penalty if $d_j - d_i + \delta_{ij}$ is greater than 0, when $\mathbf{p}_j - \mathbf{p}_i > t_p$. These two cases align well with the findings in [10] regarding optimism and pessimism of users. Therefore, we propose the following personality regularizations:

$$\sum_{\substack{(i,j),\\ \mathbf{o}_i - \mathbf{o}_j > t_o}} \max(0, d_j - d_i + \gamma_{ij})^2, \tag{3}$$

$$\sum_{\substack{(i,j),\\ \mathbf{p}_j - \mathbf{p}_i > t_p}} \max(0, d_j - d_i + \delta_{ij})^2, \tag{4}$$

where $d_i = \mathbf{U}_i\mathbf{V}\mathbf{U}^\top \cdot \mathbf{1}_{n\times 1}$. Next, we will show that by minimizing Eq. 3 and Eq. 4, we can model personality information:

- Eq. 3 imposes a penalty when the user contradicts optimism behavior, where $d_j - d_i + \gamma_{ij} > 0$ and $\mathbf{o}_i - \mathbf{o}_j > t_o$. Minimizing $d_j - d_i + \gamma_{ij}$ will force d_i to be closer to higher values and d_j to be close to lower values. Thus, u_i is more likely to establish positive links resulting in an increase in its degree while u_j is less likely to have positive links so that its degree will be decreased.
- Eq. 4 imposes a penalty when the user contradicts pessimism behavior, where $d_j - d_i + \delta_{ij} > 0$ and $\mathbf{p}_j - \mathbf{p}_i > t_p$. Minimizing $d_j - d_i + \delta_{ij}$ will force d_j to be closer to lower values and d_i to be closer to higher values. Therefore, u_j is more likely to establish negative links toward others resulting in a decrease in its degree while u_i is less likely and as a consequent its degree will increase.

The above observations show that we can model personality information by minimizing Eq.3 and Eq.4.

4.3 Personality Information for Signed Link Prediction

Having added the personality regularizations, our proposed framework *SLP* now seeks to solve the following optimization problem:

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{W} \odot (\mathbf{G} - \mathbf{U}\mathbf{V}\mathbf{U}^\top)||_{\mathbf{F}}^2 + \lambda_1||\mathbf{U}||_{\mathbf{F}}^2 + \lambda_2||\mathbf{V}||_{\mathbf{F}}^2 + \tag{5}$$
$$\alpha \sum_{\substack{(i,j),\\ \mathbf{o}_i - \mathbf{o}_j > t_o}} \max(0, d_j - d_i + \gamma_{ij})^2 + \beta \sum_{\substack{(i,j),\\ \mathbf{p}_j - \mathbf{p}_i > t_p}} \max(0, d_j - d_i + \delta_{ij})^2$$

where $d_i = \mathbf{U}_i\mathbf{V}\mathbf{U}^\top \cdot \mathbf{1}_{n\times 1}$ and $\mathbf{1}_{n\times 1} \in \mathbb{R}^{n\times 1}$ is a vector with all elements equal to 1. The additional information related to the optimism and pessimism of users alleviates the data sparsity problem for signed link prediction. If u_i does not have any positive/negative links, learning her latent factor is impossible. However, if we have information on how much u_i is optimistic or pessimistic, we still can learn $\mathbf{U}_i$ for u_i via personality regularization.

Since the optimization problem in Eq.5 is not jointly convex with respect to $\mathbf{U}$ and $\mathbf{V}$, there is no neat closed solution for it due to existence of the max function. Therefore, it can be rewritten into its matrix form as:

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{W} \odot (\mathbf{G} - \mathbf{U}\mathbf{V}\mathbf{U}^\top)||_F^2 + \lambda_1||\mathbf{U}||_{\mathbf{F}}^2 + \lambda_2||\mathbf{V}||_F^2$$

Algorithm 1 The Proposed Framework SLP

Input: $\mathbf{G}$, $\mathbf{o}$, $\mathbf{p}$, $\{\delta, \gamma, \lambda_1, \lambda_2, \alpha, \beta, d, t_o, t_p\}$.
Output: $\widetilde{\mathbf{G}}$

1: Initialize $\mathbf{U}$ and $\mathbf{V}$ randomly
2: **while** Not convergent **do**
3: Calculate $\mathbf{R}_{ij}^{\gamma,k}$ and $\mathbf{R}_{ij}^{\delta,k}$ according to Eq. 7
4: Calculate $\frac{\partial \mathcal{J}_k}{\partial \mathbf{U}}$ and $\frac{\partial \mathcal{J}_k}{\partial \mathbf{V}}$ accoording to Eq. 10
5: Update $\mathbf{U} \leftarrow \mathbf{U} - \gamma_u \frac{\partial \mathcal{J}_k}{\partial \mathbf{U}}$
6: Update $\mathbf{V} \leftarrow \mathbf{V} - \gamma_v \frac{\partial \mathcal{J}_k}{\partial \mathbf{V}}$
7: **end while**
8: Set $\widetilde{\mathbf{G}} = \mathbf{U}\mathbf{V}\mathbf{U}^\top$

$$+ \alpha||\mathbf{R}^{\gamma,k} \odot (\mathbf{D} - \mathbf{D}^T + \gamma)||_F^2 + \beta||\mathbf{R}^{\delta,k} \odot (\mathbf{D} - \mathbf{D}^T + \delta)||_F^2 \tag{6}$$

where $\mathbf{D} = \mathbf{1}_{n\times n} \cdot \mathbf{U}\mathbf{V}^\top\mathbf{U}^\top$. We define $\mathbf{R}^{\alpha,k} \in \mathbb{R}^{n\times n}$ and $\mathbf{R}^{\beta,k} \in \mathbb{R}^{n\times n}$ in the k-th iteration,

$$\mathbf{R}_{ij}^{\gamma,k} = \begin{cases} 1, & \text{if } \mathbf{o}_i - \mathbf{o}_j > t_o \text{ and } d_j + \gamma_{ij} > d_i \\ 0, & \text{otherwise} \end{cases},$$

$$\mathbf{R}_{ij}^{\delta,k} = \begin{cases} 1, & \text{if } \mathbf{p}_j - \mathbf{p}_i > t_p \text{ and } d_j + \delta_{ij} > d_i \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

Expanding the objective function of Eq. 6 in the k-th iteration can be written as,

$$\min_{\mathbf{U},\mathbf{V}} \mathcal{J}_k = Tr(\mathbf{W} \odot (-2\mathbf{G}^\top + \mathbf{U}\mathbf{V}^\top\mathbf{U}^\top)\mathbf{U}\mathbf{V}\mathbf{U}^\top) + \alpha\boldsymbol{h}(\gamma) + \beta\boldsymbol{h}(\delta) \tag{8}$$

where the function $\boldsymbol{h}(x)$ is defined as,

$$\boldsymbol{h}(x) = Tr(\mathbf{R}^{x,k} \odot \Big((\mathbf{D} - \mathbf{D}^T + x)(\mathbf{D}^T + x^\top) + (\mathbf{D}^T - \mathbf{D} - x)\mathbf{D}\Big)), \tag{9}$$

where x could be either γ or δ. We use the gradient descent method to solve Eq. 8, which has been proven to gain an efficient solution [21]. The partial derivations of $\mathcal{J}_k$ w.r.t. $\mathbf{U}$ and $\mathbf{V}$ are,

$$\begin{aligned} \frac{1}{2}\frac{\partial \mathcal{J}_k}{\partial \mathbf{U}} =& \mathbf{W} \odot ((\mathbf{U}\mathbf{V}\mathbf{U}^\top - \mathbf{G})\mathbf{U}\mathbf{V}^\top) + \mathbf{W} \odot ((\mathbf{U}\mathbf{V}^\top\mathbf{U}^\top - \mathbf{G}^\top)\mathbf{U}\mathbf{V}) + \\ & \alpha(\mathbf{I}f(\gamma)\mathbf{U}\mathbf{V} + f(\gamma)\mathbf{I}\mathbf{U}\mathbf{V}^\top) + \beta(\mathbf{I}f(\delta)\mathbf{U}\mathbf{V} + f(\delta)\mathbf{I}\mathbf{U}\mathbf{V}^\top) \\ \frac{1}{2}\frac{\partial \mathcal{J}_k}{\partial \mathbf{V}} =& \mathbf{U}^\top(\mathbf{W} \odot (\mathbf{U}\mathbf{V}\mathbf{U}^\top - \mathbf{G}))\mathbf{U} - \alpha\mathbf{U}^\top f(\gamma)\mathbf{I}\mathbf{U} - \beta\mathbf{U}^\top f(\delta)\mathbf{I}\mathbf{U} \end{aligned} \tag{10}$$

where

$$\begin{aligned} \boldsymbol{g}(x) &= (\mathbf{R}^{x,k} \odot (\mathbf{D} + x))^\top + \mathbf{R}^{x,k} \odot \mathbf{D}^\top, \\ I &= \mathbf{1}_{n\times n}, \\ f(x) &= g(x)^\top - g(x) \end{aligned} \tag{11}$$

With the partial derivations of $\mathbf{U}$ and $\mathbf{V}$, an optimal solution of the objective function in Eq. 6 can be obtained as shown in Algorithm 1.

The inputs to this algorithm are user-user positive/negative link matrix $\mathbf{G}$, expected degree difference matrices δ_α and δ_β and users' optimism $\mathbf{o}$ and pessimism $\mathbf{p}$ scores. We randomly initialize $\mathbf{U}$ and $\mathbf{V}$ in line 1. From line 2 to 8, we update $\mathbf{U}$ and $\mathbf{V}$ until they converge. The algorithm will stop when objective function has little change or when it reaches a predefined maximal iterations. The output of algorithm is the estimated positive/negative link matrix which is

computed as $\widetilde{\mathbf{G}} = \mathbf{U}\mathbf{V}\mathbf{U}^\top$. $|\widetilde{\mathbf{G}_{ij}}|$ shows the likelihood of establishing relation from u_i to u_j and $sign(\widetilde{\mathbf{G}_{ij}})$ also shows whether the relation is a positive or negative link.

4.4 Time Complexity

The most time-consuming operations are the calculations of $\frac{\partial \mathcal{J}}{\partial \mathbf{U}}$ and $\frac{\partial \mathcal{J}}{\partial \mathbf{V}}$. We only discuss the time complexity analysis of these two steps. Let the number of non-zero elements of $\mathbf{W}$ be N_w.

- We first focus on the analysis of $\mathbf{R}^{\gamma,k}$ which needs to be calculated in each iteration k and is $O(N_u nd + nd^2 + {N_t}^2)$. We can keep the difference between users' optimism scores-$|\mathbf{o}_i - \mathbf{o}_j|$ before running the algorithm. If N_t is considered the number of pair of users $|\mathbf{o}_i - \mathbf{o}_j| > t$, we only need to check whether $d_j + \gamma_{ij} > d_i$ is satisfied for these pairs or not. Due to the sparsity of $\mathbf{U}_i\mathbf{V}\mathbf{U}^\top$, the time complexity of calculating $d_i = \mathbf{U}_i\mathbf{V}\mathbf{U}^\top \cdot \mathbf{1}_{n\times 1}$ is $O(N_u nd + nd^2)$ if N_u is the number of non-zero elements of $\mathbf{U}_i\mathbf{V}\mathbf{U}^\top$. The time complexity for comparing N_t number of users is also ${N_t}^2$. The same analysis can be done for $\mathbf{R}^{\delta,k}$. Therefore, the total time complexity of calculating $\mathbf{R}^{\gamma,k}$ and $\mathbf{R}^{\delta,k}$ in each iteration is $O(N_u nd + nd^2 + {N_t}^2)$.
- For the analysis of $\frac{\partial \mathcal{J}}{\partial \mathbf{U}}$ in Eq.10, since $\mathbf{W}$ is very sparse, the time complexity of calculating both $\mathbf{W} \odot ((\mathbf{U}\mathbf{V}\mathbf{U}^\top - \mathbf{G})\mathbf{U}\mathbf{V}^\top)$ and $\mathbf{W} \odot ((\mathbf{U}\mathbf{V}^\top\mathbf{U}^\top - \mathbf{G}^\top)\mathbf{U}\mathbf{V})$ are $O(nd^2 + N_w d)$ in which N_w is the number of non-zero elements of $\mathbf{W}$. Similarly, since $\mathbf{D} + \gamma$ and $\mathbf{D} + \delta$ only need to be computed once, we can get the time complexity of $\mathbf{I}f(\gamma)\mathbf{U}\mathbf{V} + f(\gamma)\mathbf{I}\mathbf{U}\mathbf{V}^\top$ and $\mathbf{I}f(\delta)\mathbf{U}\mathbf{V} + f(\delta)\mathbf{I}\mathbf{U}\mathbf{V}^\top$ as $O(N_u nd + N_{\gamma,k} n + nd^2)$ and $O(N_u nd + N_{\delta,k} n + nd^2)$, respectively. $N_{\gamma,k}$ and $N_{\delta,k}$ denote the numbers of non-zero elements of $\mathbf{R}^{\gamma,k}$ and $\mathbf{R}^{\delta,k}$, respectively. For brevity, we omit the detailed analysis of these terms. The time complexity of $\frac{\partial \mathcal{J}}{\partial \mathbf{U}}$ in each iteration is thus $O(nd^2 + N_u nd + (N_w + N_{\delta,k} + N_{\gamma,k})n)$.
- Now we discuss the time complexity of $\frac{\partial \mathcal{J}}{\partial \mathbf{V}}$ in Eq.10. Since the number of non-zero elements of $\mathbf{W}$ is N_w, the cost of calculating $\mathbf{W} \odot (\mathbf{U}\mathbf{V}\mathbf{U}^\top - \mathbf{G})$ is $O(nd^2 + N_w d)$. Thus, the time complexity of $\mathbf{U}^\top(\mathbf{W} \odot (\mathbf{U}\mathbf{V}\mathbf{U}^\top - \mathbf{G}))\mathbf{U}$ is $O(nd^2 + N_w d)$. With the same approach for calculating the time complexities of $\mathbf{I}f(\gamma)\mathbf{U}\mathbf{V} + f(\gamma)\mathbf{I}\mathbf{U}\mathbf{V}^\top$ and $\mathbf{I}f(\delta)\mathbf{U}\mathbf{V} + f(\delta)\mathbf{I}\mathbf{U}\mathbf{V}^\top$, the time complexity of computing $\mathbf{U}^\top f(\gamma)\mathbf{I}\mathbf{U}$ and $\beta\mathbf{U}^\top f(\delta)\mathbf{I}\mathbf{U}$ is $O(N_u nd + N_{\gamma,k} n + N_{\delta,k} n + nd^2)$. Hence, the total time complexity of $\frac{\partial \mathcal{J}}{\partial \mathbf{V}}$ in each iteration is $O(nd^2 + N_u nd + (N_w + N_{\delta,k} + N_{\gamma,k})n)$.

5 EXPERIMENTS

In this section, we first introduce the datasets we use, and then explore the effectiveness of *SLP* for signed link prediction. In particular, we design an extensive set of experiments to (1) evaluate the proposed method *SLP*, (2) investigate the effect of personality information on the performance of *SLP* and (3) conduct parameter analysis to examine the sensitivity of *SLP* to the main parameters.

Table 1: Statistics of the datasets.

	Epinions	Slashdot
# of Users	131,828	7,275
# of Positive Links	717,677	67,705
# of Negative Links	123,705	20,851
# of Ratings/Opinions	577,692	403,896

5.1 Datasets

We collect two large datasets of online signed social networks, Epinions and Slashdot. We perform some standard preprocessing steps by filtering out users without either positive or negative links. Table 1 shows the statistics of our datasets.

Epinions. this is a product-review website where users can establish trust and distrust relationships. We treat each trust and distrust relation as positive and negative links and construct user-user matrix $\mathbf{G}$ where $\mathbf{G}_{ij} = 1$ if user i trusts user j, and $\mathbf{G}_{ij} = -1$ if user i distrusts user j. Also, $\mathbf{G}_{ij} = 0$ if the information is missing. Users can rate each item in a scale of 1 to 5. Here we assume $r_{th} = 3$, i.e., scores in $\{1, 2, 3\}$ are treated as low and $\{4, 5\}$ as high scores [38]. We define optimism and pessimism based on the user's rating behavior in Epinions, following the scenario 1 [10].

Slashdot. this is a technology-related news platform which allows users to tag each other as either 'friend' or 'foe'. Similar to the Epinions, we construct user-user matrix $\mathbf{G}$ from the positive (friendship) and negative links (foes) in the network. Likewise, users can express their opinions toward each other by annotating the articles posted by each other. We define a user's optimism and pessimism based on her opinion expressing behavior following scenario 2 [10].

5.2 Experimental Settings

We use 5-fold cross validation for evaluation. Each time, we hold one fold out and treat it as our test set $\mathcal{A}$. From the remaining 4 folds, we pick x% of positive and x% of negative links to construct our training set $\mathcal{T}$. Then we set $\mathbf{G}_{ij} = 0, \forall \langle u_i, u_j \rangle \in \mathcal{T} \cup \mathcal{A}$ and new representation of $\mathbf{G}$ is fed to each predicator. In this paper, we vary x as $\{50, 60, 70, 80, 90, 100\}$ to investigate how well our model performs with different sizes of training set. As stated before, in signed networks, positive links are often much denser than negative ones; hence signed links are imbalanced in both training and test sets. Therefore we rather use the Area Under Curve (AUC), to assess the performance of signed link predictors on predicted values of links between pairs of users in test set, $\{\widetilde{\mathbf{G}}_{ij}, \forall \langle u_i, u_j \rangle \in \mathcal{A}\}$.

To evaluate the proposed framework *SLP*, we compare it with the following representative signed link predictors:

- **MF** [21]: This is a variant of the proposed method and preforms matrix factorization on $\mathbf{G}$. This baseline ignores the personality regularization. We select this method to see how signed link prediction method performs in absence of personality information.
- **DB/OP/RP** [34]: This method first extracts two sets of topological-based features for each pair of users: the first set consists of seven (7) degree-based (DB) features, and the second set contains twelve (12) features describing user's optimism/reputation (OP/RP), which are derived from the links between users. A total of 19 features. Then, it trains a logistic regression

Table 2: Performance comparison on Epinions.

	50%	60%	70%	80%	90%	100%
SLP	**0.8097**	**0.8113**	**0.8193**	**0.8258**	**0.8363**	**0.8504**
MF	0.6497	0.6585	0.6614	0.6779	0.6884	0.6957
DB/OP/RP	0.5981	0.6093	0.6107	0.6202	0.6379	0.6426
All23	0.5699	0.5721	0.5740	0.5797	0.5813	0.5934
TDP	0.5461	0.5532	0.5591	0.5613	0.5826	0.6001
Random	0.4939	0.4994	0.5044	0.5024	0.5009	0.4997

Table 3: Performance comparison on Slashdot.

	50%	60%	70%	80%	90%	100%
SLP	**0.8308**	**0.8404**	**0.8464**	**0.8518**	**0.8613**	**0.8725**
MF	0.6784	0.6810	0.6911	0.7066	0.7177	0.7265
DB/OP/RP	0.6347	0.6401	0.6476	0.6558	0.6716	0.6811
All23	0.5913	0.5983	0.6005	0.6094	0.6126	0.6237
TDP	0.5710	0.5794	0.5816	0.5871	0.5981	0.6013
Random	0.4924	0.5023	0.4981	0.4993	0.5003	0.5011

classifier using these features to predict positive/negative links for a given pair.

- **All23** [27]: This method considers 23 different topological structure features for each pair of relation between users. It then trains a logistic regression classifier to predict positive/negative links. The deployed features could be categorized into two groups. The first category captures the local relations of a node to the rest of the network including positive/negative in-degrees and out-degrees. The second category is extracted according to the balance theory.
- **TDP** [19]: This baseline treats positive (trust)/negative (distrust) link propagation as a repeating sequence of atomic operations. In this propagation-based method, positive link propagates multiple steps while negative link propagates only a single step.
- **Random**: This algorithm assigns a random sign (+ or -) to each user pair.

Note we do not compare *SLP* with any traditional positive link predictor such as [24, 37]. This is because the signed link prediction problem could not be carried out by trivially applying positive link predictors [39], despite existence of several positive link predictors [24, 37]. We use cross-validation to determine the best values for the proposed method and the baselines with parameters. For the proposed framework, we set the parameters as follows: $\{\lambda_1 = \lambda_2 = 0.1,\ d = 100,\ t_p = 0.5,\ t_o = 0.5,\ \alpha = 80,\ \beta = 80\}$. We also construct the matrices δ^{α} and δ^{β} as follows,

$$\gamma_{ij} = \begin{cases} 15, & \text{if } \Delta\mathbf{R}^{\alpha}_{\mathbf{o}_{ij}} \geq 1500 \\ 10, & \text{if } \Delta\mathbf{R}^{\alpha}_{\mathbf{o}_{ij}} \geq 200 \\ 5, & \text{otherwise} \end{cases} \quad \delta_{ij} = \begin{cases} 15, & \text{if } \Delta\mathbf{R}^{\beta}_{\mathbf{p}_{ij}} \geq 500 \\ 10, & \text{if } \Delta\mathbf{R}^{\beta}_{\mathbf{p}_{ij}} \geq 150 \\ 5, & \text{otherwise} \end{cases} \tag{12}$$

where $\Delta\mathbf{R}^{\alpha}_{\mathbf{o}_{ij}}$ and $\Delta\mathbf{R}^{\beta}_{\mathbf{p}_{ij}}$ indicate the difference between the ranks of users u_i and u_j, assuming users are sorted in an descending order according to their optimism and pessimism scores, respectively.

5.3 Performance Comparison

The comparison results are shown in Table 2 and Table 3 for Epinions and Slashdot, respectively. We observe that all methods outperform the baseline Random. The performance of each method improves with the increasing size of the training data. *SLP* performs the best among all methods. Next, we discuss why *SLP* did better:

- *SLP* outperforms *TDP* since edge signs could be incorporated into the signed link prediction rather than requiring a notion of propagation from farther-off parts of the network as [19] did. Moreover, in contrast to *SLP*, *TDP* does not consider imbalance distribution of positive/negative links.
- *SLP* always outperforms *All23* since features extracted based on the topological structure, may not be robust due to the sparsity problem, and there might be many pairs of users without features based on balance theory [14]. The imbalance problem of positive/negative links distribution cannot be handled by *All23*.
- *SLP* achieves better performance over *DB/OP/RP*, despite that both of the approaches leverage optimism/reputation-based features. The reason is *DB/OP/RP* uses topological structure to extract these features and hence suffers from the sparsity problem, similar to *All23*. Simply put, there could be many pairs of users with zero optimism/reputation, which make leveraging the optimism/reputation-based features less useful in alleviating the imbalance problem of signed links distributions. In contrast, *SLP* infers users' personality information from their feedback on different issues other than merely using signed links.
- *SLP* has better performance than *MF*. This is because *SLP* incorporates personality information to predict positive/negative links while *MF* does not. This suggests the importance of personality information in the problem of signed link prediction.

We perform t-test on all comparisons and the t-test results suggest that all improvements are significant. To recap, the proposed framework obtains significant performance improvement by exploiting personality information.

5.4 Further Experiment with Personality Information

In this set of experiments, we probe further if the newly discovered personality-based features in *SLP* can be incorporated into other methods such as *All23* and *DB/OP/RP*. The addition of personality-based features to these two methods results two variants as follows.

- **All23+PI** [10, 27]: This is a variant of *All23* [27] which considers four additional personality-based features (i.e., optimism and pessimism [10]) for each pair of users in addition to the existing 23 structural based features.
- **DB/OP/RP+PI**[10, 34]: This is a variant of *DB/OP/RP* [34] which uses the same four personality-based features [10] along with the existing topological-based features.

The comparison results are shown in Table 4 and Table 5 for Epinions and Slashdot. We observe the following:

Table 4: The impact of personality information on different signed link predictors on Epinions.

	50%	60%	70%	80%	90%	100%
SLP	**0.8097**	**0.8113**	**0.8193**	**0.8258**	**0.8363**	**0.8504**
DB/OP/RP+PI	0.6761	0.6839	0.6947	0.7014	0.7177	0.7263
DB/OP/RP	0.5981	0.6093	0.6107	0.6202	0.6379	0.6426
All23+PI	0.6531	0.6673	0.6701	0.6848	0.6974	0.7012
All23	0.5699	0.5721	0.5740	0.5797	0.5813	0.5934

Table 5: The impact of personality information on different signed link predictors on Slashdot.

	50%	60%	70%	80%	90%	100%
SLP	**0.8308**	**0.8404**	**0.8464**	**0.8518**	**0.8613**	**0.8725**
DB/OP/RP+PI	0.6938	0.7076	0.7188	0.7254	0.7374	0.7513
DB/OP/RP	0.6347	0.6401	0.6476	0.6558	0.6716	0.6811
All23+PI	0.6847	0.6918	0.7043	0.7181	0.7203	0.7328
All23	0.5913	0.5983	0.6005	0.6094	0.6126	0.6237

- Both *All23+PI* and *DB/OP/RP+PI* outperform their corresponding original methods, i.e., *All23* and *DB/OP/RP*. This performance gain confirms the added value of the proposed personality features extracted from an exogenous source of information, in addition to those inferred from the signed links.
- Even with the personality-based features, *All23+PI* and *DB/OP/RP+PI* were outperformed by *SLP* because *SLP* uses latent features for signed link prediction problem, while *All23+PI* and *DB/OP/RP+PI* both use a manually developed set of features.

5.5 Impact of Personality Information on SLP

We further explore the impact of personality information on signed link prediction, in an attempt to capture the different strength levels of optimism and pessimism that manifest online. First, we assume the size of training set is fixed to 100% throughout this section. Then, using k-means, we divide the users in the training set into two groups, based on their optimism and pessimism scores: (1) $\mathcal{S}$: users with strong personality (both high optimism and pessimism), and (2) $\mathcal{I}$: indifferent users (with low optimism and pessimism).

To assess the impact of personality information, we train our model on the whole training data, in three different ways by : (1) only considering the personality information (optimism and pessimism scores) of users in $\mathcal{S}$ (i.e., $\mathbf{o}_i = 0, \mathbf{p}_i = 0, \forall u_i \in \mathcal{I}$), (2) only considering the personality of users in $\mathcal{I}$, (3) considering the personality of users in $\mathcal{S} \cup \mathcal{I}$. The results corresponding to these three runs are shown in Table 6 with the following observations:

- By removing users with strong personality, the performance of *SLP* drastically drops. This can be observed by comparing the performance on $\mathcal{I}$ and $\mathcal{S} \cup \mathcal{I}$.
- The removal of indifferent users' personality does not have significant influence on the performance of *SLP*, as we compare the performance on $\mathcal{S}$ and $\mathcal{S} \cup \mathcal{I}$.

The role of personality information becomes clearer via this study: personality information can help signed link prediction with a caveat– the stronger the personality, the more improvement in performance. In other words, low personality information as indifferent users exhibit, can lead to performance deterioration as shown in the performance on $\mathcal{I}$. Moreover, with only a slight difference, the personality information from users in $\mathcal{S} \cup \mathcal{I}$, is worthless compared to the information from $\mathcal{S}$. This supports the fact that the information from users in $\mathcal{I}$ does not make much difference. In other words, the personality information from indifferent users could be treated as irrelevant to the signed link prediction problem.

Table 6: The impact of personality information on SLP.

(a) Epinions

$\mathcal{S}$	$\mathcal{I}$	$\mathcal{S} \cup \mathcal{I}$
0.8189	0.7677	0.8504

(b) Slashdot

$\mathcal{S}$	$\mathcal{I}$	$\mathcal{S} \cup \mathcal{I}$
0.8406	0.8094	0.8725

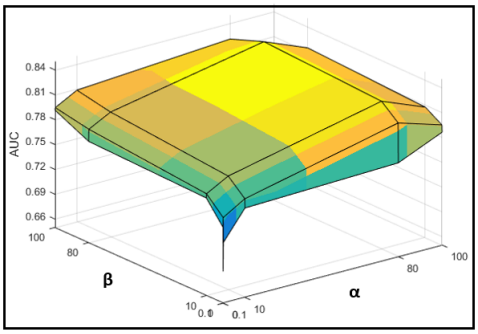
5.6 Parameter Sensitivity Analysis of SLP

SLP has two important parameters, α and β which control the contribution from user's optimism and pessimism, respectively. Here, we discuss their effect by varying both α and β as $\{0, 0.1, 10, 80, 100\}$ when the size of training set is fixed to 100%. The results are shown in Figure 1 with the followings observed.

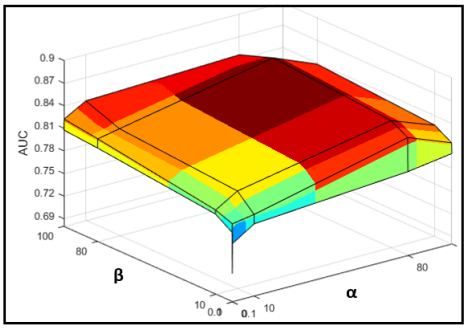
- In general, the performance of *SLP* increases with the increase of α and β, and then it degrades. These patterns ease the parameter selection for *SLP*.
- The performance improves even when α and β slightly change from 0 to 0.1, which confirms the importance of user personality information in signed link prediction.
- After reaching to certain values, continuing to increase α and β will result in the performance reduction. This suggests that large values of α and β dominate the learning process and the model could learn U and V inaccurately due to the overfitting of the model to the personality information.
- *SLP* is more sensitive to β than α since α controls the contribution of user's optimism, which is denser than pessimism information.

6 RELATED WORK

The ease of using the Internet has raised numerous security and privacy issues. Mitigating these concerns has been studied from different aspects such as identifying malicious activities [1–3], addressing users' privacy issues [5, 7, 9] and finding positive/ negative links (i.e., trust/distrust) between users [10, 11, 36, 37]. Positive link prediction (a.k.a. trust prediction) has been extensively studied [6, 24, 37] in which the goal is to predict only positive links from existing ones. The availability of signed networks has motivated the research on signed link prediction [14, 27]. The recent advances on signed link prediction demonstrate that negative links have added value in addition to positive links [39]. Thus, we focus on the positive/negative link prediction problem which is different from positive link prediction [37] and sign prediction [25, 42]. It is more challenging compared to them because: first, in contrast to positive link prediction, in signed link prediction, we aim to predict both positive and negative links simultaneously and second, the



(a) Epinions



(b) Slashdot

Figure 1: Effect of personality regularization in SLP.

sign prediction problem only infers the signs of the existing links while, in the signed link prediction, we predict the existence of a link between a pair of nodes as well as its sign.

The existing signed link prediction studies can be divided into two categories: supervised and unsupervised methods. Supervised methods consider the signed link creation problem as a classification problem by using the old links and training a classifier on the features extracted from the signed networks[14, 27]. For example, the work of [27] first extracts in-degree and out-degree from signed links and then uses balance and status theory to extract triangle-based features and it then verifies the importance of balance and status theory for signed link prediction. Another work of [14] extends the triangle-based features to the k-cycle-based features.

Unsupervised methods often perform predictions based on certain topological properties of signed networks [21, 35, 43]. One type of unsupervised methods is node similarity based methods [35], which first define similarity metrics to calculate node similarities, and then provide a way to predict the signed relations based on them. Another type of unsupervised methods is propagation-based methods [16, 19, 44]. Positive sign propagation is treated as a repeating sequence of matrix operations [19]. Negative sign propagation then stopped after multiple steps of positive sign propagation [19]. The work of [16] considers ignorance as well as partial positive/negative links in its proposed positive/negative link propagation method by modeling the network as an intuitive fuzzy relations. Another category of unsupervised methods is based on low-rank matrix factorization [21, 40, 43]. The work in [21] models the signed link prediction problem as a low-rank matrix factorization model, based on the weak structural balance on the signed network. Another study in [43] extends the low-rank model to perform link prediction across multiple signed networks. Also the work of Wang et al. [40] completes a binary matrix with positive and negative elements in an online setting by penalizing the difference between predicted matrix and the ground truth by logistic loss and matrix max-norm. Authors of [38] incorporate side information, i.e., helpfulness ratings in a low-rank matrix factorization model to predict negative links. Beigi et al. [11] incorporates users' emotional information in a low-rank matrix factorization framework to predict positive and negative links between them. Another work of [36] predicts the negative links using positive links and content-centric user interactions.

Likewise, *SLP* is also based on the low-rank matrix factorization model. The difference between *SLP* and the above low-rank matrix factorization models is that we investigate the role of personality information, i.e. optimism/pessimism in signed link prediction. Since optimists tend to establish more positive links than others and pessimists establish more negative links [17], we model this fact as a constraint in the low-rank matrix factorization cost function to guide the learning process of $\mathbf{U}$ and $\mathbf{V}$.

Exploiting user's features such as trustworthiness, bias, and optimism has been discussed in prior studies [30, 34]. For example, the work of [34] addresses the problem of sign prediction based on users' optimism/reputation. The authors define optimism as users' voting pattern towards others and the reputation as their popularity. Specifically, their approach calculates the optimism as the difference between the number of user's positive and negative out-links. Moreover, they introduce rank based optimism and reputation based on the rank of users in the signed social network. Another work [30] computes bias and prestige of nodes based on the positive links between users in signed social networks. It defines bias as the user's truthfulness. The prestige is also calculated based on the opinion of other users in the form of inlinks a user gets.

Similarly, we calculate users' optimism/pessimism and exploit these information for signed link prediction. The difference between our work and the above works is that we infer users' optimism/pessimism based on a source other than signed links, i.e. users' feedback and interactions on different entities (items and posts). These additional sources of information could help to overcome the data sparsity and imbalance problem for signed link data.

7 CONCLUSION AND FUTURE WORK

In this paper, we study the role of user personality, in particular, optimism and pessimism to mitigate the data sparsity problem in signed link prediction. User's feedback is further used to estimate optimism/pessimism. We then investigate the incorporation of such information for predicting positive/negative links and solving data sparsity problem for signed link prediction. We propose the framework *SLP* by mathematically incorporating optimism and pessimism information. We evaluate *SLP* on real-world datasets and the results demonstrate the complementary role of personality information in the signed link prediction problem. This work

also enables us to experiment if incorporating different types of personality information can be relevant or helpful.

One future direction is to consider additional user's personality traits including extraversion, agreeableness, openness, conscientiousness and neuroticism as suggested in the Big Five Model [28]. We also plan to explore more on inferring users personal information by incorporating content information. Furthermore, we can expand our statistic solution of user behavior to a dynamic one as users can evolve over time and they adapt to different situations, though slowly. A nuanced solution is to consider temporal dynamics of user's personality for dynamic signed link prediction. Another future direction is to explore the impact of users' personality and signed links on recommendation systems as signed links have previously shown promising results for recommendation task.

ACKNOWLEDGMENTS

This material is based upon the work supported, in part, by NSF #1614576, ARO W911NF-15-1-0328 and ONR N00014-17-1-2605.

REFERENCES

[1] Hamidreza Alvari, Elham Shaabani, and Paulo Shakarian. 2018. Early Identification of Pathogenic Social Media Accounts. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 169–174.

[2] Hamidreza Alvari and Paulo Shakarian. 2019. Hawkes Process for Understanding the Influence of Pathogenic Social Media Accounts. In *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*. IEEE.

[3] Hamidreza Alvari, Paulo Shakarian, and JE Kelly Snyder. 2017. Semi-supervised learning for detecting human trafficking. *Security Informatics* 6, 1 (2017), 1.

[4] Jens B Asendorpf and Susanne Wilpers. 1998. Personality effects on social relationships. *Journal of personality and social psychology* 74, 6 (1998), 1531.

[5] Ghazaleh Beigi, Ruocheng Guo, Alexander Nou, Yanchao Zhang, and Huan Liu. 2019. Protecting User Privacy: An Approach for Untraceable Web Browsing History and Unambiguous User Profiles. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 213–221.

[6] Ghazaleh Beigi, Mahdi Jalili, Hamidreza Alvari, and Gita Sukthankar. 2014. Leveraging community detection for accurate trust prediction. (2014).

[7] Ghazaleh Beigi and Huan Liu. 2018. Privacy in social media: Identification, mitigation and applications. *arXiv preprint arXiv:1808.02191* (2018).

[8] Ghazaleh Beigi and Huan Liu. 2018. Similar but different: Exploiting usersâĂŹ congruity for recommendation systems. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 129–140.

[9] Ghazaleh Beigi, Kai Shu, Yanchao Zhang, and Huan Liu. 2018. Securing social media user data: An adversarial approach. In *Proceedings of the 29th on Hypertext and Social Media*. ACM, 165–173.

[10] Ghazaleh Beigi, Jiliang Tang, and Huan Liu. 2016. Signed Link Analysis in Social Media Networks. In *Tenth International AAAI Conference on Web and Social Media*.

[11] Ghazaleh Beigi, Jiliang Tang, Suhang Wang, and Huan Liu. 2016. Exploiting emotional information for trust/distrust prediction. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 81–89.

[12] Ronald S Burt, Joseph E Jannotta, and James T Mahoney. 1998. Personality correlates of structural holes. *Social Networks* 20, 1 (1998), 63–87.

[13] Edward C Chang. 1998. Distinguishing between optimism and pessimism: A second look at the optimism–neuroticism hypothesis.. In *International Congress of Psychology*. American Psychological Association.

[14] Kai-Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S Dhillon. 2011. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of international conference on Information and knowledge management*.

[15] Jinsook Cho. 2006. The mechanism of trust and distrust formation and their relational outcomes. *Journal of retailing* 82, 1 (2006), 25–35.

[16] Martine De Cock and Paulo Pinheiro Da Silva. 2005. A many valued representation and propagation of trust and distrust. In *International Workshop on Fuzzy Logic and Applications*. 114–120.

[17] Andrew L Geers, Sean P Reilley, and William N Dember. 1998. Optimism, pessimism, and friendship. *Current Psychology* 17, 1 (1998), 3–19.

[18] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI extended abstracts on human factors in computing systems*.

[19] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2004. Propagation of trust and distrust. In *Proceedings of WWW*.

[20] Russell Hardin. 2004. Distrust: Manifestations and management. In *Distrust*.

[21] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S Dhillon. 2012. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD*.

[22] Rong Hu and Pearl Pu. 2013. Exploring Relations between Personality and User Rating Behaviors.. In *UMAP Workshops*.

[23] Rong Hu and Pearl Pu. 2014. Exploring personality's effect on users' rating behavior. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*.

[24] Min-Hee Jang, Christos Faloutsos, and Sang-Wook Kim. 2014. Trust prediction using positive, implicit, and negative information. In *Proceedings of WWW*.

[25] Amin Javari and Mahdi Jalili. 2014. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links. *ACM TIST* (2014).

[26] Deborah Welch Larson and Russell Hardin. 2004. Distrust: Prudent, if not always wise. *Distrust* 8 (2004), 34.

[27] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In *Proceedings of WWW*.

[28] Robert R McCrae and Oliver P John. 1998. An introduction to the five-factor model and its applications. *Personality: critical concepts in psychology* 60 (1998).

[29] D Harrison McKnight and Norman L Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies*. Springer, 27–54.

[30] Abhinav Mishra and Arnab Bhattacharya. 2011. Finding the bias and prestige of nodes in networks based on trust scores. In *Proceedings of WWW*.

[31] Michael F Scheier and Charles S Carver. 1985. Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health psychology* 4, 3 (1985).

[32] Michael F Scheier, Charles S Carver, and Michael W Bridges. 1994. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. *Journal of personality and social psychology* 67, 6 (1994), 1063.

[33] Michael F Scheier, Charles S Carver, and Michael W Bridges. 2001. Optimism, pessimism, and psychological well-being. *Optimism and pessimism: Implications for theory, research, and practice* 1 (2001), 189–216.

[34] Mohsen Shahriari, Omid Askari Sichani, Joobin Gharibshah, and Mahdi Jalili. 2016. Sign prediction in social networks based on users reputation and optimism. *SNAM* (2016).

[35] Panagiotis Symeonidis and Nikolaos Mantas. 2013. Spectral clustering for link prediction in social networks with positive and negative links. *Social Network Analysis and Mining* 3, 4 (2013), 1433–1447.

[36] Jiliang Tang, Shiyu Chang, Charu Aggarwal, and Huan Liu. 2015. Negative link prediction in social media. In *Proceedings of WSDM*. ACM, 87–96.

[37] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Exploiting homophily effect for trust prediction. In *Proceedings of WSDM*. ACM, 53–62.

[38] Jiliang Tang, Xia Hu, Yi Chang, and Huan Liu. 2014. Predictability of distrust with interaction data. In *Proceedings of CIKM*. ACM, 181–190.

[39] Jiliang Tang, Xia Hu, and Huan Liu. 2014. Is distrust the negation of trust?: the value of distrust in social media. In *Proceedings of HT*.

[40] Jing Wang, Jie Shen, Ping Li, and Huan Xu. 2017. Online Matrix Completion for Signed Link Prediction. In *Proceedings of WSDM*. ACM.

[41] Xin Xin, Irwin King, Hongbo Deng, and Michael R Lyu. 2009. A social recommendation framework based on multi-scale continuous conditional random fields. In *Proceedings of conference on Information and knowledge management*.

[42] Shuang-Hong Yang, Alexander J Smola, Bo Long, Hongyuan Zha, and Yi Chang. 2012. Friend or frenemy?: predicting signed ties in social networks. In *SIGIR*.

[43] Jihang Ye, Hong Cheng, Zhe Zhu, and Minghua Chen. 2013. Predicting positive and negative links in signed social networks by transfer learning. In *WWW*.

[44] Cai-Nicolas Ziegler and Georg Lausen. 2005. Propagation models for trust and distrust in social networks. *Information Systems Frontiers* 7, 4-5 (2005), 337–358.

Extracting Meaningful Insights on City and Zone Levels Utilizing US Open Government Data

Samaa Gazzaz[†]
Computer Science Department
King Abdulaziz University, Jeddah,
Saudi Arabia, sgazzaz@kau.edu.sa

Praveen Rao
Department of Computer Science and Electrical Engineering
University of Missouri-Kansas City,
Kansas City, Missouri, USA,
raopr@umkc.edu

ABSTRACT

It is estimated that merely 4% of the world's population reside on US soil. Remarkably, 43% of the entire population of prominent websites are hosted in the United States (Fig. 1). Even though most data content on the Web is unstructured, the US government has had big contributions in producing and actively releasing structured datasets related to different fields such as health, education, safety and finance.

Aforementioned datasets are referred to as Open Government Data (OGD) and are aimed at increasing the structured data pool in conjunction with promoting government transparency and accountability. In this paper, we present a new system "OGDXplor" which processes raw OGD through a well-defined procedure leveraging machine learning algorithms and produces meaningful insights.

The novelty of this work is encompassed by the collective approach utilized in developing the system and tackling challenges. First by addressing arising challenges due to data being collected and aggregated from heterogeneous sources that otherwise would have been impossible to acquire as a comprehensive unit. moreover, classification and comparisons are drawn on a much finer level that we refer to as zone level. Zones are the areas encompassed and defined by zip codes and are seldomly used in classifying and extracting insights as presented here. OGDXplor facilitates comparing and classifying zones located in different cities or zones within an individual city.

The system is presented to end-users as a web application allowing users to elect zones and features relevant to their use case. Results are presented in both chart and map formats which aids the decision-making process.

CCS CONCEPTS

• **Information systems~Clustering** • **Information systems~Web applications** • **Information systems~Clustering and classification** • Information systems~Data cleaning • Mathematics of computing~Dimensionality reduction • Mathematics of computing~Cluster analysis • Software and its engineering~Process validation • Software and its engineering~Use cases

KEYWORDS

Data science, open data, clustering, feature selection.

ACM Reference format:

Samaa Gazzaz and Praveen Rao. 2019. Extracting Meaningful Insights on City and Zone Levels Utilizing US Open Government Data. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316470

1 Introduction

Publishing Open Government Data (OGD) has been a growing subject of interest among governments within the last decade. The US is considered the largest influential country in producing data on the Web (followed by Germany hosting only 8%)[14]. It is estimated that the use of OGD when developing applications and services can yield $3 trillion in income across global economy [16]. This would yield better decision-making, trend-recognition and prediction of future events [16].

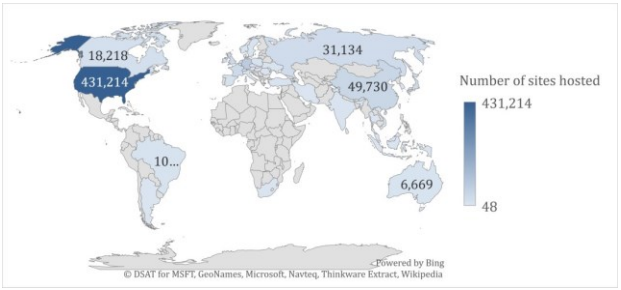


Figure 1: Top 100 web hosting countries with respect to the top 1 million influential websites.

[†]This work was done while the author was affiliated with the University of Missouri-Kansas City.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316470

Providing access to OGD promotes the involvement of citizens within their low and high-level governments. In addition, it affirms the transparency and accountability of said governments. Nonetheless, when it comes to OGD, the datasets are usually burdened with weaknesses whether in regards to its completion, accuracy or conforming to a unified form of publication. The release of such datasets does not guarantee the availability of means to interpret, visualize and analyze this data. Although federal and state governments within the United States contribute extensively to the body of open data [20], the datasets remain inadequate in facilitating critical decision-making processes. Moreover, due to the existence of various data sources and the absence of a regulating body for open government data, accuracy and conformation challenges arise.

In this paper, we present a new approach exploiting OGD and machine learning algorithms aiming at producing a user-friendly application OGDXplor enabling users to gain insights from the datasets. First, data is collected from multiple sources and dataset weaknesses are addressed. Next, irrelevant features are eliminated, and data is aggregated and clustered based on relevant features. Finally, the system features include clustering zones and cities based on relevant features acquired from feature selection. Results are presented to the user in easy-to-follow formats aiding decision making and insight extraction. Those insights will enable users to learn more about different zones and cities comparing data related to health, education, safety, and more.

The paper is organized as follows, in section 2 we discuss the related work and the different challenges faced when dealing with OGD. In section 3, our approach in tackling those challenges, system flow and methods used are discussed. In section 4, evaluation of OGDXplor's accuracy and results of the system are presented. Finally, in section 5, we conclude the paper.

2 Related Work and Challenges

2.1 Related Work

Even though the open data concept is relatively immature [20], there has been an abundant number of research applications based on open data sources. Moreover, the movement towards utilizing OGD expanded when hundreds of national and local governments started releasing OGD portals [20]. In [21], researchers discuss utilizing "open-access satellite data" in the field of biodiversity research. Open data is transformed by applying different techniques and extracting meaningful information from raw input. In addition, numerous research opted to gathering datasets from a variety of commercial "non-government" sources optimizing the benefits of analysis and visualization of data [10][12]. Nonetheless, most of the literature is fixated at explaining the open data initiative, its advantages and disadvantages, and how beneficial it could be if adapted in the right manner [9][6][8][16][2]. On the other hand, we seldom come across a system that is built on heterogeneous OGD gathered from diverse government agencies and structured into a meaningful system.

One project to be highlighted is Data USA [5] in which researchers gathered OGD and implemented a visualization system for extracting facts about areas in the US. Data USA utilizes a collection of datasets from varying government sources to create one comprehensive website that delivers a user-friendly application where the use of the data is optimized [1].

Although it introduced a solution to the existing problem with multi-source open datasets [8], Data USA does not provide pattern recognition in similarities between multiple cities/zones, future possible occurrences, and recommended actions for decision-makers. In OGDXplor on the other hand, we introduce a system model that utilizes both heterogeneous OGD and machine learning techniques in extracting meaningful insights, recognizing patterns among cities and zones and facilitating the decision-making process.

This submission version of your paper should not have headers or footers, Baldassare (2000). It should remain in a one-column format—please do not alter any of the styles or margins.

2.2 Challenges

OGD portals offer huge potential when it comes to insightful understanding of the trends behind the data enabling an informed decision process. Unfortunately, while obtaining and processing raw OGD, several challenges arise with respect to the data collection process, understanding the meaning of the data in question, and processing heterogeneous data through the same pipeline. These challenges are summarized in figure 2.

In the following sections, we focus on discussing the most commonly faced challenges and their implications.

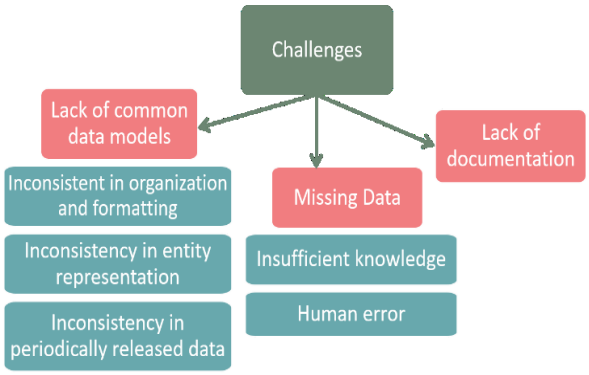


Figure 2: Challenges associated with collecting OGD especially from heterogeneous sources.

2.2.1 Lack of Common Data Models.:One of the most occurring challenges is the lack of common data models. As a result of the multi-source data collection process, data models are recognized to be very inconsistent from source to source. For example, when collecting crime data released by Los Angeles city and Chicago city data portals, we identify the inconsistency in organization and formatting. Data is labeled differently and organized based on different features and properties. Both datasets cannot be combined seamlessly without manual alteration.

Moreover, inconsistency in entity representation is another common challenge. We can look at the crime datasets from both Kansas City and Chicago to immediately recognize that inconsistency. Even though both datasets are concerned with public safety and crime information, we notice the vast difference of entity (represented by a row) interpretation in those datasets. In the Chicago crime data, each entity involves information about the case number, primary type, data and description of a crime. In this case, that information implies that each entity represents a crime. On the other hand, in Opendata KC, each entity is described by information such as involvement, race, sex and age, which in return implies that each entity represents a person involved with a crime (whether a suspect or a victim).

Finally, each source has different release period for each dataset where, for example, one source could release data every year grouped by month while the other releases every quarter grouped by location.

*2.2.2 Missing Data.:*Another frequently occurring problem is missing data. Whether it is the result of insufficient knowledge about a specific feature or simply human error when entering data, missing information does not only impair the full understanding of the information provided, but also hinders the ability to infer and predict future trends in an unbiased fashion. Insufficient information can occur while generating or entering the data. For example, there are cases where zip code information was not attainable when entering data resulting in entries such as 99999, 00000 or XXXXX as the zip code value. In addition, incomplete knowledge when the data was being generated results in leaving out attribute values that appear as missing data in published dataset.

3 Our Approach

Initially, datasets are collected from heterogeneous sources such as local governments and privately-owned businesses. In order to address datasets' issues, it is crucial to first handle each dataset separately. During which, we address missing data and the abundance of irrelevant features using feature extraction. Afterwards, datasets are collectively aggregated and merged to generate a coherent dataset that describes a wide range of features. Finally, clustering the cities and zipcode areas (i.e. zones) based on similarities depending on features that can be specified by end users. An overview of OGDXplor is visualized in figure 3. In this section, we discuss in depth, the process and the approach.

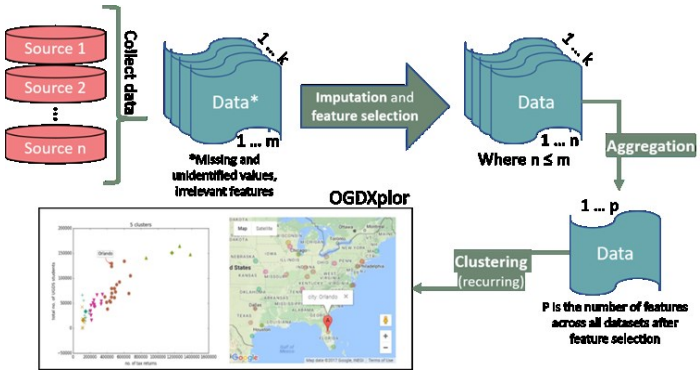


Figure 3: OGDXplor system overview.

3.1 Data Collection and Preparation

In order to establish a solid foundation of the system, datasets are collected from OGD portals with good documentation and consistency in release periods. These collected datasets (Table 1) play a great role providing the system with feature that will be useful to end users. Datasets collected from the department of energy (https://energy.gov/) mainly contained information regarding the rates of utility companies within proximity of a city or a zip code. Those included investor and non-investor owned companies in addition to the service type and commercial/industrial/residential rates.

Moreover, datasets collected from the department of education (https://www.ed.gov/) were comprehensive nationwide statistics about colleges and universities in the US. That information includes more than 7800 colleges and encompasses more than 40 attributes regarding each college. From the department of agriculture (https://www.agriculture.gov/), we collected datasets which included information regarding areas and nearby farmer's markets nationally. This information can be an important factor in many decisions such as area to live or start a local produce market or restaurant.

The department of treasury (https://www.treasury.gov/) provides valuable information regarding taxes filed by tax payers nationwide. This information includes counts of all individually/joint filed taxes, number of dependents, in addition to other data all mapped to zip code areas in the US. Finally, datasets collected from the department of defense (https://www.defense.gov/) included information about the residency of military personnel within the US. This dataset also provides information about age/gender/racial demographics of the enlisted recruits. All collected data is in tabular form. Data is transformed into JSON format for its flexibility preparing for aggregation based on zip codes.

Table 1: Data sources and brief description of content.

Data source	Time period	Number of attributes	Number of entries	Brief Description of content
Department of Energy	2014-2015	9	34k+	Information regarding rates of utility companies
Department of Education	2013-2015	40+	7800+	Extensive statistics about nationwide colleges
Department of Agriculture	2013	2	440k+	Listing of areas and nearby farmer's market
Department of Treasury	2013	100+	27k+	Taxes filed nationwide and filing information
Department of Defense	2010	30+	940+	Defense military recruits enlisted

3.2 Feature Selection

Feature selection is defined as the election of the attributes that most closely represent the whole dataset fairly, even when other attributes are missing. Usually, feature selection is used for dimensionality reduction and pattern recognition in a dataset distribution [22].

The most prominent technique for dimensionality reduction is Principal Component Analysis (PCA), where the resultant features are the outcome of the mapping to the lower level space [22]. On the other hand, our intentions in this application are different since we aim to select a subset of the existing features rather than find a mapping to a new lower dimension. Principal Feature Analysis (PFA) [22] is an adaptation of PCA that allows the retention of previously existing features even after the reduction of dimensionality. As the first step of PFA, the covariance matrix is calculated from the original dataset such that each entry in the resulting matrix is defined as follows:

$$\rho_{ij} = \frac{E[x_i x_j]}{E[x_i^2]E[x_j^2]}$$

Next, we compute the principal components as in PCA and the eigenvalues of the covariance matrix. The retained variability must be established before choosing the subspace dimension. Then, we cluster the data using K-means and use the Euclidean distance to decide where each data point resides. Finally, for each cluster, obtain the corresponding feature that closely represent that cluster and consider this feature as a Principal Feature. The resulting is a list of the most relevant features.

3.3 Data Imputation

There is extensive research in the area of data imputation, and we can categorize data imputation techniques to: mean substitution, regression and K-Nearest Neighbor imputation. In mean substitution, we calculate the mean of all the values in the same feature and impute the result value in all missing cells. This technique is the fastest, but it imposes risk of introducing bias. Regression imputation utilizes the trend analysis of existing values and predicts the missing value based on the trend. This technique becomes expensive as the size of the dataset increases. In addition, it is mostly used to impute datasets that are missing values in a single feature. K-Nearest Neighbor (KNN) technique only considers k entities out of the whole dataset in imputing the missing value. Those k entities are usually chosen based on similarity to the entity with the missing value. Next, the values in the k entities are averaged, resulting in the imputed value.

For the purposes of this research, we utilize KNN as it does not introduce the kind of bias that mean substitution introduces, nor is computationally expensive as regression. KNN algorithm can be generally used in multiple applications such as estimation, classification and imputation [18]. In the case of imputation, the choice of the number of nearest neighbors to consider is very critical. As a rule of thumb, it is preferred to consider $k = \sqrt{n}$ where n is the number of entities in the dataset [18]. Considering $\sqrt{n}$ entities as the nearest neighbors to reference when imputing missing data ensures that we only consider entities that are similar to the entity whose missing field we are trying to impute.

3.4 Clustering

Since the beginning, our goal was to deliver a system that enables users to compare and differentiate cities and zones upon features of their selection. In order to provide that ability to distinguish between the different areas, clustering is utilized where areas are grouped based on similarity. Clustering is perfect for our dataset since it is used as part of unsupervised learning. In order to cluster, we need to select the "optimal" number of clusters desired. Choosing the optimal k is a broad research area where multiple techniques have been developed. The most famous yet is the Gap statistic [19]. In this approach, they utilize the within-cluster dispersion to decide the estimated number of clusters from a clustering algorithm's results [19]. Where Dr is the sum of all data points in a cluster and Wk is the within-cluster sum of squares around the center of the cluster, we calculate Gap statistic as:

$$D_r = \sum_{i,i' \in C_r} d_{ii'}, \qquad W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

We show how the Gap statistic optimizes the number of clusters in the following example. To cluster zones in El Paso, Texas, based on two features: number of tax returns and number of dependents, we first calculate the Gap statistic. Estimating the number of clusters k over k = 2, 3, 4, 5, and 6, the result is shown on the left in figure 4 where the Gap value is high when k=3. As sown on the right in figure 4, clustering over 3 groups gives a clear boundary to each group of zones within El Paso.

After figuring out the best value for k for a specific configuration, we start the clustering process via k-means algorithm. In here, we employ the Lloyd's algorithm which implements k-means iteratively to converges to local minimum in lowest amount of time:

$$C_k = \{x_n : ||x_n - \mu_k|| \leq all\ ||x_n - \mu_l||\}$$

$$\mu_k = \frac{1}{C_k} \sum_{x_n \in C_k} x_n$$

The notation denotes that each cluster C_k is a set of points x_n such that the distance from a mean is minimized. The symbol μ_k represents the mean of cluster k.

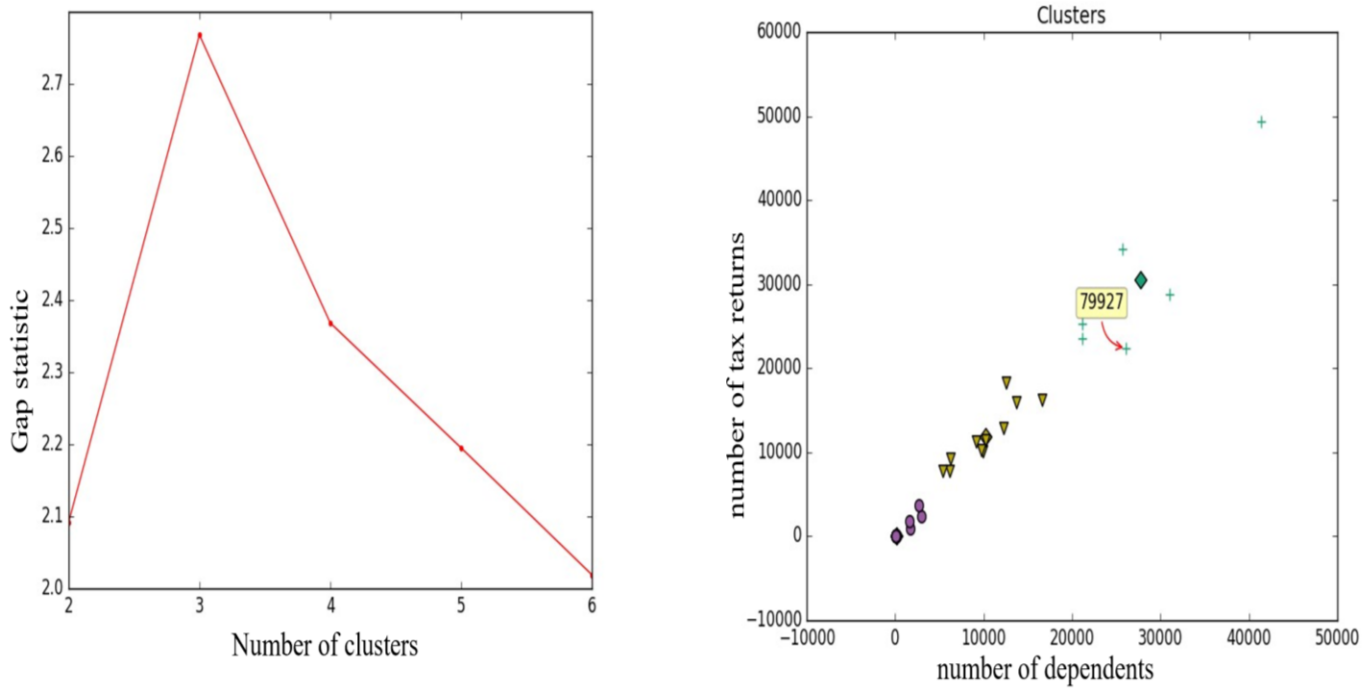


Figure 4: Gap statistics results for k = 2, 3, 4, 5, 6 (left) and Sample date clustered based on number of tax returns and dependents (right).

4 Evaluation and Results

Cluster analysis is a broad science concerned with evaluating clustering technique and evaluating the optimal number of clusters. When evaluating clustering validity, three validation criteria can be explored. External criteria, which consider a pre-specified structure when evaluating outcomes of a clustering algorithm [17]. Relative criteria, which evaluate a clustering algorithm's results by comparting them to results from other clustering algorithms [17]. Internal criteria, evaluating the outcomes of a clustering based on a calculated value involving entities in the dataset within the evaluation process [17]. We will utilize internal criteria for cluster validity.

For internal criteria, there are two main features that are considered when validating: compactness and separation. Compactness refers to ensuring the minimization of the distance among data points within a single cluster (e.g. variance can be used to calculate compactness) [3]. Separation criteria favors higher distances between cluster centers (i.e. distinct cluster assignments) [3]. We can calculate the separation among two clusters by measuring the distance between: the closest data points, the furthest data points, or the centers of the two clusters. This is referred to as single linkage, complete linkage, and comparison of centroids, respectively [3]. In order to evaluate internal criteria, multiple validation indices were introduced to evaluate the compactness and separation levels of clusters.

In order to evaluate our clustering approach, involving the use of the Gap statistic as input to k-means clustering, we compare the recommended number of clusters provided by the Gap statistic with other internal criteria-validation indices. The goal of this evaluation is to detect how accurate the clustering is when the number of clusters is determined by the Gap statistic. This is done by comparing the estimated number of clusters with the results from the following indices: Silhouette index [15] [13], Calinski-Harabasz index [11] [4], Dunn index [11] [7] and Davis-Bouldin index [11] [13].

The experiment is divided into two categories: clustering over features elected by feature selection, and clustering over randomly chosen features. In the first, features are the most relevant representation of data; while in the second, selection is simulating user activity. The hope is to recognize that our approach performs well under both circumstances.

The results of the experiments are shown in table 2 and figure 5. In the figure, results of validating clustering over randomly selected features are indicated in orange and feature selection results are in blue. Values presented are averaged out based on cross validation. Data was clustered into 2, 3, 4, 5, and 6 clusters respectively, where each time the value of the index is calculated. The results shown indicate a pattern of conformation among the indices. For example, when features are randomly

collected, the recommended number of clusters by 3 of the 4 indices is 5 clusters. This value is the same value resulting from utilizing the Gap index in our approach. On the other hand, there is a slight misalignment between the validity metrics recommendation of optimal number of clusters when clustering using features selected by feature selection and the Gap statistic recommendation. In this case, the value recommended by the Gap statistic is 3 where none of the other indices recommend that value; however, the values recommended are close. Considering the results of our experiments, we conclude that the clustering technique and approach of the system is performing efficiently and yielding sufficient clustering.

Table 2: Validation of clustering using internal validity indices vs Gap statistic.

Validity metrics	Feature selection		Features selected randomly	
	Number of clusters	Best value	Number of clusters	Best value
Silhouette index	4	0.927	5	0.819
Calinski-Harabasz index	6	2099.5	6	2328.2
Dunn index	4	0.345	5	0.168
Davis-Bouldin index	4	0.344	5	0.161
Gap statistic	3	2.775	5	1.407

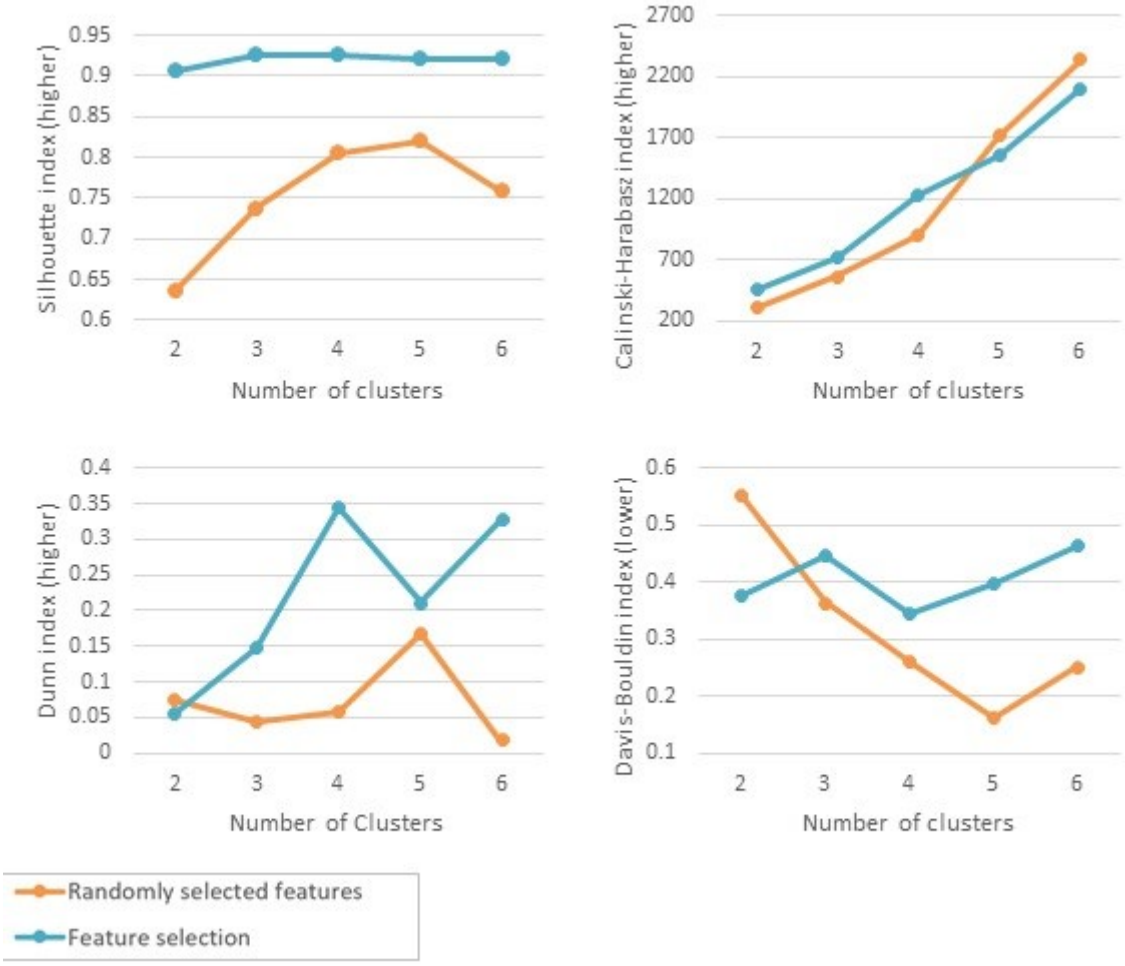


Figure 5: Cluster validation results using Silhouette index, Dunn index, Davis-Bouldin index and Calinski-Harbaz index.

4.1 Application Use Case Example

A prospective college student is interested in attending a college in one of the largest 100 cities in the US but would like to know the ratio of college students to the general populations in these cities. Moreover, they would be interested in knowing how Orlando, FL ranks in that comparison and what are similar cities to Orlando considering these properties.

Through OGDXplor, the student is able to provide the city name then specify the features she is interested in. Next, the system clusters the cities over the selected features and produce these two views to the user (shown in figure 6).

In this case, the system identified five clusters considering the total number of undergraduate students and the number of tax returns (representing the general population). In figure 6 (left), we see that the city of Orlando is located somewhat in the middle cluster while Miami is located in the upper-right cluster.

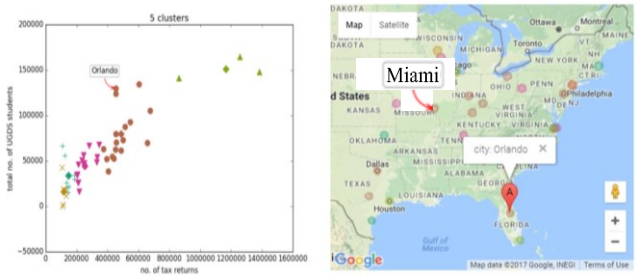


Figure 6: System result when clustering over number of tax returns and total number of undergraduate students.

5 Conclusion

In this paper we presented OGDXplor, a system that utilizes open government data and machine learning in producing visualization of the cities and zones within the US. Regardless of the dataset's weaknesses, we were able to tackle the challenges and chosen zip code information to be the distinctive key to each area. Feature selection enables the retention of the most relevant features, thus saving time and ensuring relevant results to the user. Finally, clustering was a big part of our approach in addition to utilizing Gap statistic to estimate the best possible number of clusters. In the end, the user can choose features they want to compare, choose the area and view charts and the map showing the grouping and comparing results. This system is helpful for multiple applications and of great help in the decision-making process.

REFERENCES

[1] Ann Perrin and César Hidalgo. 2016. Data USA: The Most Comprehensive Visualizations of U.S. Public Data. (April 2016). Retrieved April 3, 2017 from https://www.media.mit.edu/sponsorship/getting-value/collaborations/datausa

[2] Barbara Ubaldi. 2013. Open government data: towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance 22. OECD Publishing, Paris, France.

[3] Gordon Linoff and Michael Berry. 2011. Data Mining Techniques For Marketing, Sales and Customer Support. (3rd. ed.). John Wiley & Sons, Inc., USA.

[4] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3.1 (1974), 1-27.

[5] Data USA. 2016. Retrieved from https://datausa.io

[6] Felipe Gonzalez-Zapata and Richard Heeks. 2015. The Multiple Meanings of Open Government Data: Understanding Different Stakeholders and Their Perspectives. Government Information Quarterly 32, 4 (October 2015), 441–452.

[7] James Bezdek and Nikhil Pal. 1995. Cluster validation with generalized Dunn's indices. Proceedings of the Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, Dunedin, 1995, pp. 190-193.

[8] James Hendler, Jeanne Holm, Chris Musialek, and George Thomas. 2012. US government linked open data: semantic.data.gov. IEEE Intelligent Systems 27, 3 (May 2012), 25-31.

[9] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. Government Information Quarterly 32, 4 (October 2015), 399–418.

[10] Klaus Ackermann, Eduardo Reyes, Sue He, Thomas Keller, Paul van der Boor, and Romana Khan. 2016. Designing policy recommendations to reduce home abandonment in Mexico. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 13-20.

[11] Ujjwal Maulik and Sanghamitra Bandyopadhyay. 2002. Performance evaluation of some clustering Aagorithms and validity indices." IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 12, (December 2002), 1650–1654.

[12] Muhammad R. Khan and Joshua E. Blumenstock. 2016. Predictors without borders: behavioral modeling of product adoption in three developing countries. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 145-154.

[13] Petrovic, Slobodan. 2006. A comparison between the silhouette index and the Davies-Bouldin index in labelling ids clusters. Proceedings of the 11th Nordic Workshop on Secure IT Systems (NordSec 06). Sweden. (Oct 20, 2006).

[14] Pingdom. 2012. The US Hosts 43% of the World's Top 1 Million Websites. (July 2012). Retrieved April 2, 2017 from http://royal.pingdom.com/2012/07/02/united-stateshosts-43-percent-worlds-top-1-million-websites/

[15] Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, (1987), 53–65.

[16] Sharon S. Dawes, Lyudmila Vidiasova, and Olga Parkhimovich. 2016. Planning and designing open government data programs: an ecosystem approach. Government Information Quarterly 33, 1 (January 2016), 15–27.

[17] Sergios Theodoridis and Konstantinos Koutroumbas. 2008. Pattern Recognition (4th ed.). Academic Press.

[18] Saravanan Thirumuruganathan. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. (May 2010). Retrieved July 19, 2017 from https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introductionto-k-nearest-neighbor-knn-algorithm/

[19] Robert Tibshirani, Guenther Walther and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 2, (November 2000), 411–423.

[20] Tim Davies. 2013. Open Data Barometer - 2013 Global Report. (October 31, 2013), 24-35. Open Data Institute, World Wide Web Foundation. Retrieved April 3, 2017 from http://www.cocoaconnect.org/sites/default/files/publication/Open-Data-Barometer-2013Global-Report.pdf

[21] Woody Turner, Carlo Rondinini, Nathalie Pettorelli, Brice Mora, Allison Leidner, Zoltan Szantoi, Graeme Buchanan, Stefan Dech, John Dwyer, Martin Herold, Lian Koh, Peter Leimgruber, Hannes Taubenboeck, Martin Wegmann, Martin Wikelski, and Curtis Woodcock. 2015. Free and open-Access satellite data are key to biodiversity conservation. Biological Conservation 182 (February 2015), 173–176.

[22] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian. 2007. Feature selection using principal feature analysis. In Proceedings of the 15th ACM international conference on Multimedia (MM '07). ACM, New York, NY, USA, 301-304.

Lightning Talk - Ruuh: A Conversational Social Agent

Meghana Joshi
Microsoft, India
mejoshi@microsoft.com

Sonam Damani
Microsoft, India
sodamani@microsoft.com

Khyatti Gupta
Microsoft, India
khgupt@microsoft.com

Nitya Raviprakash
Microsoft, India
niravipr@microsoft.com

ABSTRACT

Dialogue systems and conversational agents are becoming increasingly popular in the modern society but building an agent capable of holding intelligent conversation with its users is a challenging problem for artificial intelligence. In this talk, we share challenges and learnings from our journey of building a deep learning based conversational social agent called "Ruuh" (m.me/Ruuh) developed by a team at Microsoft India to converse on a wide range of topics. The authors are co-creators of Ruuh and the original paper was presented in NeurIPS 2018 Demonstration Track by two of the authors. As a social agent, Ruuh needs to think beyond the utilitarian notion of merely generating "relevant" responses and meet a wider range of user social needs. The agent also needs to detect and respond to abusive language, sensitive topics and trolling behavior of users. Some of the above objectives pose significant research challenges in the areas of NLP, IR and AI. Our agent has interacted with over 2 million real world users till date which has generated over 150 million user conversations. We intend to walk the audience through our journey of overcoming several research challenges to become the most popular social agent in India.

ACM Reference Format:
Meghana Joshi, Sonam Damani, Khyatti Gupta, and Nitya Raviprakash. 2019. Lightning Talk - Ruuh: A Conversational Social Agent. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316472

1 TECHNICAL DETAILS

In this section we will discuss a brief overview of the technical aspects of our approach.

Retrieving relevant responses

We model the task of providing relevant chat responses as an Information Retrieval problem, where for a given user message M and context C, the system retrieves and ranks the candidates by relevance and outputs one of the highest scoring responses. We create an index of paired tweets and their responses offline. At runtime, we first use TF-IDF-based fetch to generate a candidate set, and then extract features using a convolutional deep structured semantic network, followed by a ranker trained to select the best response. [3].

Emotion detection

As humans, on reading "Why don't you ever text me!", we can either interpret it as a sad or an angry emotion and the same ambiguity exists for machines too. Lack of facial expressions and voice modulations make detecting emotions in text a challenging problem. However, the success of our agent depends on our ability to modulate responses based on user emotions. In [2], we propose a Deep Learning based approach to detect emotions in textual dialogues. The essence of our approach lies in combining both semantic and sentiment-based representations for more accurate emotion detection.

Detecting offensive conversations

Users often abuse and provoke the agents to elicit inappropriate or controversial responses. We actively identify "controversial topics" and make clever dodging techniques to avoid responding to such topics. This problem is wrought with challenges like handling natural language ambiguity, rampant use of spelling mistakes and variations for offensive terms, and disambiguation with entity names such as pop songs which usually have abusive terms in them. We use a neural Bi-directional LSTM based model for this task [4].

Human-like image commenting

Users often interact with social agents by sharing their personal pictures, other images and videos. In such scenarios, agents are not expected to routinely describe the facts within the image but to express some interesting emotions and opinions about it. We use a modified version of [1], where the model is learnt using millions of image-comment pairs mined from social network websites like Instagram, Twitter etc.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316472

REFERENCES

[1] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[2] Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *Neu-IR: The SIGIR 2017 Workshop on Neural Information Retrieval* (2017).

[3] Abhay Prakash, Chris Brockett, and Puneet Agrawal. 2016. Emulating human conversations using convolutional neural network-based IR. *Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval* (2016).

[4] Harish Yenala, Manoj Chinnakotla, and Jay Goyal. 2017. Convolutional Bi-directional LSTM for Detecting Inappropriate Query Suggestions in Web Search. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 3–16.

Lightning Talk - Think Outside the Dataset: Finding Fraudulent Reviews using Cross-Dataset Analysis

Shirin Nilizadeh
University of Texas, at Arlington
shirin.nilizadeh@uta.edu

Hojjat Aghakhani
University of California Santa Barbara
hojjat@ucsb.edu

Eric Gustafson
University of California Santa Barbara
edg@ucsb.edu

Christopher Kruegel
University of California Santa Barbara
chris@ucsb.edu

Giovanni Vigna
University of California Santa Barbara
vigna@ucsb.edu

ABSTRACT

Many crowd-sourced review platforms, such as Yelp, TripAdvisor, and Foursquare, have sprung up to provide a shared space for people to write reviews and rate local businesses. With the substantial impact of businesses' online ratings on their selling [2], many businesses add themselves to multiple websites to more easily be discovered. Some might also engage in reputation management, which could range from rewarding their customers for a favorable review, or a complex review campaign, where armies of accounts post reviews to influence a business' average review score.

Most of previous work use supervised machine learning, and only focus on textual and stylometry features [1, 3, 4, 7]. Their obtained ground truth data is not large and comprehensive [4–8, 10]. These works also assume a limited threat model, *e.g.,* an adversary's activity is assumed to be found near sudden shifts in the data [8], or focused on positive campaigns.

We propose OneReview , a system for finding fraudulent content on a crowd-sourced review site, leveraging correlations with other independent review sites, and the use of textual and contextual features. We assume that an attacker may not be able to exert the same influence over a business' reputation on several websites, due to increased cost. OneReview focuses on isolating anomalous changes in a business' reputation across multiple review sites, to locate malicious activity without relying on specific patterns. Our intuition is that a business's reputation should not be very different in multiple review sites; *e.g.,* if a restaurant changes its chef or manager, then the impact of these changes should appear on reviews across all the websites. OneReview utilizes *Change Point Analysis* method on the reviews of every business independently on every website, and then uses our proposed *Change Point Analyzer* to evaluate change-points, detect those that do not match across the websites, and identify them as suspicious. Then, it uses supervised machine learning, utilizing a combination of textual and metadata features to locate fraudulent reviews among the suspicious reviews.

We evaluated our approach, using data from two reviewing websites, Yelp and TripAdvisor, to find fraudulent activity on Yelp. We obtained Yelp reviews, through the Yelp Data Challenge [9], and used our Change Point Analyzer to correlate this with data crawled from TripAdvisor. Since realistic and varied ground truth data is not currently available, we used a combination of our change point analysis and crowd-labeling to create a set of 5,655 labeled reviews. We used k-cross validation (k=5) on our ground truth and obtained 97% (+/- 0.01) accuracy, 91% (+/- 0.03) precision and 90% (+/- 0.06) recall. The model was used on the suspicious reviews, which classified 61,983 reviews, about 8% of all reviews, as fraudulent.

We further detected fraudulent campaigns that are actively initiated by or targeted toward specific businesses. We identified 3,980 businesses with fraudulent reviews, as well as, 14,910 suspected spam, where at least 40% of their reviews are classified as fraudulent. We also used community detection algorithms to locate several large astroturfing campaigns. These results show the effectiveness of OneReview in detecting fraudulent campaigns.

CCS CONCEPTS

• **Information systems → Trust**; **Reputation systems**.

KEYWORDS

Fraudulent Reviews; Cross-Dataset Change-Point Analysis;

ACM Reference Format:
Shirin Nilizadeh, Hojjat Aghakhani, Eric Gustafson, Christopher Kruegel, and Giovanni Vigna. 2019. Lightning Talk - Think Outside the Dataset: Finding Fraudulent Reviews using Cross-Dataset Analysis. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316477

REFERENCES

[1] Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, and Giovanni Vigna. 2018. Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE.
[2] Michael Anderson and Jeremy Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122, 563 (2012), 957–989.
[3] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.
[4] Nitin Jindal and Bing Liu. 2008. Opinion Spam and Analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. ACM, New York, NY, USA, 219–230. https://doi.org/10.1145/1341531.1341560
[5] Jiwei Li, Myle Ott, Claire Cardie, and Eduard H Hovy. 2014. Towards a General Rule for Identifying Deceptive Opinion Spam.. In *ACL (1)*. Citeseer, 1566–1576.
[6] Yuming Lin, Tao Zhu, Hao Wu, Jingwei Zhang, Xiaoling Wang, and Aoying Zhou. 2014. Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 261–264.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316477

[7] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 309–319.
[8] Mahmudur Rahman, Bogdan Carbunar, Jaime Ballesteros, George Burri, Duen Horng, et al. 2014. Turning the Tide: Curbing Deceptive Yelp Behaviors.. In *SDM*. SIAM, SIAM, 244–252.
[9] Yelp. 2016. Yelp Dataset Challenge. https://www.yelp.com/dataset_challenge.
[10] Kyung-Hyan Yoo and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. *Information and communication technologies in tourism 2009*.

Lightning Talk–Towards Robust Detection of Cyberbullying in Social Media

Mengfan Yao
University at Albany, State University of New York
Department of Computer Science
myao@albany

Charalampos Chelmis
University at Albany, State University of New York
Department of Computer Science
cchelmis@albany.edu

Daphney–Stavroula Zois
University at Albany, State University of New York
Electrical and Computer Engineering Department
dzois@albany.edu

ABSTRACT

The potentially detrimental effects of cyberbullying have led to the development of numerous automated, data–driven approaches, with emphasis on classification accuracy. Cyberbullying, as a form of abusive online behavior, although not well–defined, is a repetitive process, i.e., a sequence of aggressive messages sent from a bully to a victim over a period of time with the intent to harm the victim.

Existing work has focused on harassment (i.e., using profanity to classify toxic comments independently) as an indicator of cyberbullying, disregarding the repetitive nature of this harassing process. However, raising a cyberbullying alert immediately after an aggressive comment is detected can lead to a high number of false positives. At the same time, two key practical challenges remain unaddressed: (i) timeliness: the state–of–the–art relies on a fixed set of features learned during training for offline detection (i.e., after all correspondence has become available), hindering the ability to respond in a timely manner (i.e., as soon as possible) to cyberbullying events. (ii) scalabilty: the scalability of existing methods to the staggering rates at which content is generated (e.g., 95 million photos and videos are shared on Instagram per day[1]) has largely remained unaddressed.

In my lightning talk, I will introduce *CONcISE*, a novel approach for timely and accurate Cyberbullying detectiON on Instagram media SEssions, that has been accepted for presentation at the main conference [1]. Specifically, I will present a novel two–stage online approach (illustrated in Figure 1) designed to reduce the time to raise a cyberbullying alert by (i) sequentially examining comments as they become available over time, and (ii) minimizing the number of feature evaluations necessary for a decision to be made for each comment. By formalizing the problem as a sequential hypothesis testing problem, a novel algorithm has been developed that satisfies four key properties: *accuracy*, *repetitiveness*, *timeliness*, and *efficiency*.

Extensive experiments on a real–world Instagram dataset with $\sim 4M$ users and $\sim 10M$ comments demonstrate the effectiveness of the proposed approach with respect to accuracy, timeliness, efficiency, and robustness, and show that it consistently outperforms the state–of–the–art, often by a considerable margin.

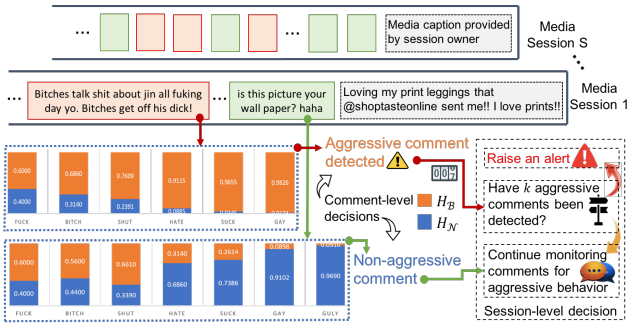


Figure 1: Overview of the proposed approach. Given a set of media objects along with their corresponding captions and hashtags, and set of comments, and an alert threshold, CONcISE examines comments as they become available over time and raises an alert only after the number of comment–level detections surpasses the threshold. The posterior probability evolution of an aggressive (upper) and non–aggressive (lower) comment as more features are examined is provided for illustration purposes. Note that the number of features used to make a decision in each case differs.

CCS CONCEPTS

• **Information systems** → *Collaborative and social computing systems and tools*; *Social networking sites*; *Data mining*; *World Wide Web*; *Social networks*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Classification; cyberharassment; optimization; sequential selection; social networks

ACM Reference Format:
Mengfan Yao, Charalampos Chelmis, and Daphney–Stavroula Zois. 2019. Lightning Talk–Towards Robust Detection of Cyberbullying in Social Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316474

[1]33 Mind–Boggling Instagram Stats & Facts for 2018: https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316474

REFERENCES

[1] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. International World Wide Web Conferences Steering Committee, 7 pages. https://doi.org/10.1145/3308558.3313462

Fake News Detection: An Interdisciplinary Research

Xinyi Zhou
Data Lab, EECS Department
Syracuse University
zhouxinyi@data.syr.edu

Reza Zafarani
Data Lab, EECS Department
Syracuse University
reza@data.syr.edu

ABSTRACT

The explosive growth of fake news and its erosion to democracy, journalism and economy has increased the demand for fake news detection. To achieve efficient and explainable fake news detection, an interdisciplinary approach is required, relying on scientific contributions from various disciplines, e.g., social sciences, engineering, among others. Here, we illustrate how such multidisciplinary contributions can help detect fake news by improving feature engineering, or by providing well-justified machine learning models. We demonstrate how news content, news propagation patterns, and users' engagements with news can help detect fake news.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing theory, concepts and paradigms.*

KEYWORDS

Fake news detection; fake news research; misinformation; disinformation; false news

ACM Reference Format:
Xinyi Zhou and Reza Zafarani. 2019. Fake News Detection: An Interdisciplinary Research. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316476

OVERVIEW

Fake news is now viewed as one of the greatest threats to democracy, journalism and economies [4]. Recent studies have shown remarkable progress on detecting fake news [1, 3]. However, interdisciplinary research on fake news detection, which can facilitate new detection approaches and provide higher interpretability, is still in its early stages. Here, we detail how contributions in various scientific disciplines can be incorporated into a machine learning framework to detect fake news. Such detection can be achieved by analyzing (1) the writing style of fake news, (2) how it propagates among users, and (3) the intentions of users spreading it.

Style-based fake news detection. It assumes that writing style of news content can help distinguish fake news from the truth. Writing style has been represented comprehensively using machine learning features at various language levels [1], or even using latent features [3]. Nevertheless, without supportive theories, the engineered/latent features can be difficult to understand and can provide limited insights into fake style. In our recent survey [4], we detail theories such as *Undeutsch hypothesis* [2] that indicate how fake news differs in content quality, sentiment, among other attributes from the truth. Such theories enable one to represent news content using interpretable *theory-driven features* that can help detect fake news. This process also facilitates further research, especially on the relationships between clickbaits and fake news [5].

Network-based fake news detection. Similarly, theories that explain how news propagates among a network of users [4], i.e., *network-based patterns* of news propagation, is an information that has not been comprehensively utilized for fake news detection. Fake news patterns in social networks can refer to the (1) news being spread, (2) spreaders of the news and (3) relationships among the spreaders. Social psychological theories provide an in-depth understanding on why such patterns exist. These patterns can then be represented at various levels of a news propagation networks (e.g., node-level, ego-level, triad-level, community-level, or whole-network-level) as features and incorporated into a supervised learning framework to detect fake news. Our initial experiments conducted on real-world data demonstrate that this approach can be highly accurate and can facilitate *fake news early detection* [4].

Intention-based fake news detection. Social psychological theories suggest that *social influence* (e.g., how widely a news article has been spread) and *self-influence* (i.e., a user's preexisting knowledge) attract regular users to fake news. Greater societal- and self-influence skew regular users more towards trusting fake news and unintentionally spreading it. This phenomenon raises a new problem: *how can we determine one's intent when posting news articles?* While the problem is yet to be explored, its solution is essential for assessing user credibility, which in turn can be embedded to detect fake news. Clearly, users that have unintentionally posted fake news should be assigned higher credibility scores compared to those who have maliciously done so. Given the limited number and scale of current public fake news datasets [5], we envision a semi-supervised learning framework to be able to measure user intentions when posting news articles and detect fake news.

REFERENCES

[1] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. *arXiv preprint arXiv:1708.07104* (2017).
[2] Udo Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der psychologie* 11 (1967), 26–181.
[3] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.
[4] Xinyi Zhou and Reza Zafarani. 2018. Fake News: A Survey of Research, Detection Methods, and Opportunities. *arXiv preprint arXiv:1812.00315* (2018).
[5] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake News: Fundamental Theories, Detection Strategies and Challenges. In *The 12th International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3289600.3291382

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316476

The Practice of Labeling: Everything You Always Wanted to Know About Labeling (But Were Afraid to Ask)

Omar Alonso
Microsoft
omalonso@microsoft.com

ABSTRACT

Many data intensive applications that use machine learning or artificial intelligence techniques depend on humans providing the initial dataset, enabling algorithms to process the rest or for other humans to evaluate the performance of such algorithms. There are, however, practical issues with the adoption of human computation and crowdsourcing at scale in the real world. Building systems data processing pipelines that require crowd computing remains difficult. In this tutorial, we present practical considerations for designing and implementing tasks that require the use of humans and machines in combination with the goal of producing high quality labels.

ACM Reference Format:
Omar Alonso. 2019. The Practice of Labeling: Everything You Always Wanted to Know About Labeling (But Were Afraid to Ask). In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3320083

1 INTRODUCTION

Many data intensive applications that use machine learning techniques depend on humans providing the initial dataset, so algorithms can process the rest or to evaluate the performance of such algorithms. Researchers and engineers depend on human computation tasks implemented in crowdsourcing platforms like Amazon Mechanical Turk or Figure Eight (aka CrowdFlower) as a scalable mechanism to produce labeled data sets. With these new platforms, not only can labeled data for training and evaluation be collected faster, cheaper, and easier than ever before, but we now see the emergence of software that combines computations performed by humans and machines in conjunction. Information retrieval, machine translation, and natural language processing (to name a few) are areas that rely heavily on such labels for research and development.

To gather useful results, a successful crowdsourced task depends on many different elements: clear instructions, representative high-quality datasets, appropriate inter-rater agreement metrics, checks on work quality, channels for worker feedback, and solid programming skills. Furthermore, designing and implementing tasks or experiments that produce thousands or millions of labels is different than conducting small scale research investigations. With the explosion of new data sources like microblogging or social networks, labeling becomes even harder, yet more crucial than ever.

In this tutorial, we present a perspective on human computation and crowdsourcing with an emphasis on practical problems and scalability. The topics and techniques described in this tutorial have a focus on designing and implementing tasks into a crowdsourcing platform and not in a specific crowdsourcing platform. That is, our goal is to describe methods that can be used independently of a specific platform or tool.

The topic of quality control, a fundamental aspect for ensuring good results, is presented from different angles as we developed tasks and not restrict the content to an overview of certain algorithms. We present examples of different data pipelines and applications that include crowdsourcing as a component for solving specific problems. We also highlight specific items that require attention when implementing large scale systems.

Engineers and researchers can expect a solid foundation for applying crowdsourcing to their own research problems or practical tasks, including labeling datasets, evaluation, and survey-style studies. Our approach to crowdsourcing not only looks at designing work from the standpoint of results, but also explicitly acknowledges the human aspects of human computation, and addresses issues that include worker motivation, fatigue, subjective judgment, and making the most of worker characteristics and diversity.

The topic of human computation has evolved significantly in the last decade with a dedicated conference, Human Computation (HCOMP), and many tutorials and workshops in top conferences. The goal of this tutorial is to identify the main concepts and techniques that practitioners should consider when asked to implement crowd-based tasks at scale.

Designing and implementing good human intelligent tasks is a solid foundation for developing real-world systems that incorporate computations performed by humans and machines working together.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320083

Crowdsourcing Inclusivity: Dealing with Diversity of Opinions, Perspectives and Ambiguity in Annotated Data

The CrowdTruth Tutorial

Lora Aroyo
Google
New York, New York, USA
l.m.aroyo@gmail.com

Anca Dumitrache
Vrije Universiteit Amsterdam
Netherlands
anca.dumitrache@vu.nl

Oana Inel
Vrije Universiteit Amsterdam
Netherlands
oana.inel@vu.nl

Zoltán Szlávik
IBM Benelux CAS
Amsterdam, Netherlands
zoltan.szlavik@nl.ibm.com

Benjamin Timmermans
IBM Benelux CAS
Amsterdam, Netherlands
benjamin.timmermans@nl.ibm.com

Chris Welty
Google Research
New York, New York, USA
cawelty@gmail.com

ABSTRACT

In this tutorial, we introduce a novel crowdsourcing methodology called *CrowdTruth* [1, 9]. The central characteristic of CrowdTruth is harnessing the diversity in human interpretation to capture the wide range of opinions and perspectives, and thus provide more reliable, realistic and inclusive real-world annotated data for training and evaluating machine learning components. Unlike other methods, we do not discard dissenting votes, but incorporate them into a richer and more continuous representation of truth. CrowdTruth is a widely used crowdsourcing methodology[1] adopted by industrial partners and public organizations such as Google, IBM, New York Times, Cleveland Clinic, Crowdynews, Sound and Vision archive, Rijksmuseum, and in a multitude of domains such as AI, news, medicine, social media, cultural heritage, and social sciences. The goal of this tutorial is to introduce the audience to a *novel approach to crowdsourcing* that takes advantage of the diversity of opinions and perspectives that is inherent to the Web, as methods that deal with disagreement and diversity in crowdsourcing have become increasingly popular. Creating this more *complex notion of truth* contributes directly to the larger discussion on *how to make the Web more reliable, diverse and inclusive.*

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Information extraction*; *Language resources*; *Lexical semantics.*

KEYWORDS

Crowdsourcing; Ambiguity; Perspectives; Diversity; Inter-annotator Disagreement; Ground Truth; Computational Social Sciences; Digital Humanities; Medical Text Annotation

[1] http://crowdtruth.org

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3320096

ACM Reference Format:
Lora Aroyo, Anca Dumitrache, Oana Inel, Zoltán Szlávik, Benjamin Timmermans, and Chris Welty. 2019. Crowdsourcing Inclusivity: Dealing with Diversity of Opinions, Perspectives and Ambiguity in Annotated Data: The CrowdTruth Tutorial. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308558.3320096

1 INTRODUCTION

The goal of the CrowdTruth tutorial is to introduce the audience to *a novel approach to crowdsourcing that takes advantage of the diversity of opinions and perspectives that is inherent to the Web.* It is relevant to the conference themes, as methods *dealing with disagreement and diversity in crowdsourcing* are increasingly popular [2, 3]; and a more complex notion of truth contributes directly to the discussion on *how to make the Web more reliable, diverse and inclusive.*

In this tutorial, we discuss how inter-annotator disagreement appears in crowdsourcing as a result of task design, ambiguous data, and crowd annotators with varying degrees of skills and reliability. We discuss elements of visual design and user interaction encouraging diversity of opinion, as well as data science methods for aggregating and interpreting crowdsourced data that accounts for inter-annotator disagreement.

The CrowdTruth methodology [1, 8] has proven to be a valuable method for gathering ground truth data for training and evaluating information extraction methods, as well as other machine learning models that facilitate data navigation and populating the Web, in a wide array of industrial and research applications at Google, IBM, Rijksmuseum, Sound and Vision, New York Times etc. These cover also a wide range of domains, from medical domain [5], to open domain [6], to cultural heritage [4], digital humanities and computational social sciences [11], to various information retrieval [12] and natural language processing tasks [7, 10].

2 DURATION AND SESSIONS

This is a hands-on tutorial with real-world examples where we discuss challenges, limitations, opportunities and open issues in an interactive manner. We provide a guided hands-on experience

with Jupyter notebooks and actual crowdsourcing templates in FigureEight[2], to enhance the learning process during the tutorial.

The tutorial material consists of presentation slides, hands-on material (Jupyter notebooks[3] and crowdsourcing templates) and relevant bibliography and is available prior, during and after the tutorial on Github[4] (CC-BY). The hands-on exercises are prepared with data gathered from experiments in real-world use cases, e.g. medical text, news text, news video and social media reviews.

The program contains four sessions dedicated to different parts of the CrowdTruth methodology:

- **Introduction:** Here we introduce crowdsourcing for gathering semantic interpretation. We describe various crowdsourcing modalities, settings and platforms in order to provide an overview of existing capabilities. Furthermore, we focus on explaining the limitations and myths of the current practices to gathering ground truth through crowdsourcing.
- **Task Design:** Here we explain how to design an open-ended or a closed task in order to gather the full range of opinions and perspectives, as well as the present ambiguity in the data. We provide a variety of examples for different data modalities such as text, image and video.
- **Data Processing:** In the CrowdTruth methodology the crowd contributor annotations are translated into a vector space representation, which is called the annotation vector. In this session, we provide examples of different annotation vectors for both open-ended and closed tasks and explain how the data collected is processed into a vector space, where the opinions and perspectives are projected, and what kind of data transformations can be applied to optimally represent in this space different data types and modalities.
- **Disagreement-aware Metrics:** We present the CrowdTruth approach for crowdsourcing ground truth data that uses disagreement-aware metrics to capture the ambiguity and the diversity of opinions and perspectives inherent in semantic interpretation. Here we focus on the quality metrics to evaluate the media units, the annotations and the crowd workers. The hands-on material provide a closer look at the data and methodology presented.

3 PREVIOUS EDITIONS AND ORGANIZATION

The tutorial organizers are long-term collaborators in the context of the **CrowdTruth development and research**[5]. The first edition of this tutorial was presented at the International Semantic Web Conference 2018[6], Monterey, California. The material and the schedule of the tutorial can be found here[7]. In the second edition, we bring together most illustrative application examples and lessons learned with practical experiences in data collection and analysis. Since 2014 different versions of this tutorial have been presented:

- as part of the **MSc curriculum** for computer science, artificial intelligence, business and information science[8] as well as in the **Digital Humanities minor** at the VU University Amsterdam. In 2016, the course was awarded the **ICT-Project of the year in education** in the Netherlands.
- in **master classes for professionals**[9] at the VU University Amsterdam and IBM Benelux,
- as a core module in the Big Data in Society Summer School held at the VU University Amsterdam, in 2015[10] and 2016[11].
- at the **Data Science with Humans in the Loop symposium**[12], organized by the Human Computation Community in the Netherlands (HComp-NL).

4 AUDIENCE

The tutorial is intended for a broad range of participants - researchers and practitioners in different domains of computer science, social science and humanities - interested in the topics of crowdsourcing, data annotation and ground truth creation. It is suitable for all levels of knowledge in human computation. Participants that are already experienced with crowdsourcing can apply CrowdTruth on their own crowdsourced datasets.

REFERENCES

[1] Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1. https://doi.org/10.15346/hc.v1i1.3

[2] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[3] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. , 11–20 pages. http://eprints.whiterose.ac.uk/122865/ © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org).

[4] Victor De Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco Van Staveren, Werner Helmich, and Dennis De Beurs. 2015. DIVE in the Event-Based Browsing of Linked Historical Media. *Web Semantics: Science, Services and Agents on WWW* 35 (2015), 152–158. http://www.websemanticsjournal.org/index.php/ps/article/view/427

[5] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Trans. Interact. Intell. Syst., Special Issue on Human-Centered Machine Learning (in publication)* 8, 2 (2017), 12. http://arxiv.org/abs/1701.02185

[6] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. False positive and cross-relation signals in distant supervision data. *arXiv preprint arXiv:1711.05186* (2017).

[7] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[8] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. (2018). https://arxiv.org/abs/1808.06080

[9] Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, and Lora Aroyo. 2017. Empirical Methodology for Crowdsourcing Ground Truth. *Semantic Web Journal, Special Issue on Human Computation and Crowdsourcing (in review)* (2017). http://www.semantic-web-journal.net/content/empirical-methodology-crowdsourcing-ground-truth-0

[10] Oana Inel and Lora Aroyo. 2017. Harnessing diversity in crowds and machines for better ner performance. In *European Semantic Web Conference*. Springer, 289–304.

[11] Oana Inel, Tommaso Caselli, and Lora Aroyo. 2016. Crowdsourcing Salient Information from News and Tweets.. In *LREC*. European Language Resources Association (ELRA), 3959–3966.

[12] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, ZoltÃąn Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. 2018. Studying Topical Relevance with Evidence-based Crowdsourcing. In *CIKM*. ACM, 1253–1262.

[2] https://www.figure-eight.com
[3] http://data.crowdtruth.org
[4] https://github.com/CrowdTruth/CrowdTruth-core
[5] http://crowdtruth.org/team/
[6] http://iswc2018.semanticweb.org
[7] crowdtruth.org/tutorial
[8] http://crowdtruth.org/course/watson-innovation-course-2016/
[9] http://btimmermans.com/2017/10/05/ibm-watson-masterclasses-with-vu-amsterdam-and-tu-delft/
[10] http://crowdtruth.org/course/big-data-summerschool-2015/
[11] http://crowdtruth.org/course/big-data-summerschool-2016/
[12] http://amsterdamdatascience.nl/event/data-science-with-humans-in-the-loop/

Designing Equitable Algorithms for the Web

Ricardo Baeza-Yates
NTENT & Northeastern University at SV, CA, USA,
rbaeza@acm.org

Sharad Goel
Stanford University
Stanford, CA, USA,
scgoel@stanford.edu

ABSTRACT

Machine learning algorithms increasingly affect both our online and offline experiences. Researchers and policymakers, however, have rightfully raised concerns that these systems might inadvertently exacerbate societal biases. We provide an introduction to fair machine learning, beginning with a general overview of algorithmic fairness, and then discussing these issues specifically in the context of the Web.

To measure and mitigate potential bias from machine learning systems, there has recently been an explosion of competing mathematical definitions of what it means for an algorithm to be fair. Unfortunately, as we show, many of the most prominent definitions of fairness suffer from subtle shortcomings that can lead to serious adverse consequences when used as an objective. To illustrate these complications, we draw on a variety of classical and modern ideas from statistics, economics, and legal theory.

We further discuss the equity of machine learning algorithms in the specific context of the Web, focusing on search engines and e-commerce websites. We expose the different sources for bias on the Web and how they impact fairness. They include not only data bias, but also biases that are produced by data sampling, the algorithms per-se, user interaction and feedback loops that result from user personalization and content creation. All these lead to a vicious cycle that affects everybody.

The content of this tutorial is mainly based in the work of the authors [1,2,3,4].

CCS Concepts/ACM Classifiers

• Computing methodologies → Machine learning.

• Information systems → Electronic commerce

• Information systems → Web Search Engines

Author Keywords

Fairness, bias, search and recommender systems.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314193

ACM Reference format:

Ricardo Baeza-Yates and Sharad Goel. 2019. Designing Equitable Algorithms for the Web. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 1 pages. https://doi.org/10.1145/3308560.3314193

1 BIOGRAPHIES

Ricardo Baeza-Yates is, since June 2016, CTO of NTENT, a semantic search technology company based in California, USA. He is also the Director of Data Science Programs at Northeastern University, Silicon Valley campus, since August 2017. Before, he was VP of Research at Yahoo Labs, based in Barcelona, Spain, and later in Sunnyvale, California, from January 2006 to February 2016. He is co-author of the best-seller Modern Information Retrieval textbook published by Addison-Wesley in 1999 and 2011 (2nd edition), that won the ASIST 2012 Book of the Year award. From 2002 to 2004 he was elected to the Board of Governors of the IEEE Computer Society and between 2012 and 2016 he was elected for the ACM Council. Since 2010 is a founding member of the Chilean Academy of Engineering. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow, among other awards and distinctions. He obtained a Ph.D. in CS from the University of Waterloo, Canada, in 1989, and his areas of expertise are web search and data mining, information retrieval, data science, and algorithms in general.

Sharad Goel is an Assistant Professor at Stanford University in the Department of Management Science & Engineering, and is the founder and director of the Stanford Computational Policy Lab. He also holds courtesy appointments in Computer Science, Sociology, and the Law School. In his research, Sharad looks at public policy through the lens of computer science, bringing a new, computational perspective to a diverse range of contemporary social issues, including police practices, voter fraud, and media bias. Before joining the Stanford faculty, Sharad completed a Ph.D. in Applied Mathematics at Cornell University, and worked as a Senior Researcher at Yahoo and Microsoft in New York City.

REFERENCES

[1] Ricardo Baeza-Yates. Bias in the Web. 2017 Gödel Lecture, Wien, June 2017. URL: https://www.youtube.com/watch?-v=CguWbbQeruM
[2] Ricardo Baeza-Yates. Bias on the Web. Communications of ACM, June 2018. URL: https://cacm.acm.org/magazines/-2018/6/228035-bias-on-the-web/fulltext
[3] Sam Corbett-Davies and Sharad Goel. Defining and Designing Fair Algorithms. Tutorial at ICML 2018, Stockholm, Sweden, July 2108.
[4] Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023v2, August 2018. URL: https://5harad.com/papers/fair-ml.pdf

Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned

Sarah Bird
Microsoft
slbird@microsoft.com

Ben Hutchinson
Google
benhutch@google.com

Krishnaram Kenthapadi
LinkedIn
kkenthapadi@linkedin.com

Emre Kıcıman
Microsoft
emrek@microsoft.com

Margaret Mitchell
Google
mmitchellai@google.com

ABSTRACT

Researchers and practitioners from different disciplines have highlighted the ethical and legal challenges posed by the use of machine learned models and data-driven systems, and the potential for such systems to discriminate against certain population groups, due to biases in algorithmic decision-making systems. This tutorial aims to present an overview of algorithmic bias / discrimination issues observed over the last few years and the lessons learned, key regulations and laws, and evolution of techniques for achieving fairness in machine learning systems. We will motivate the need for adopting a "fairness-first" approach (as opposed to viewing algorithmic bias / fairness considerations as an afterthought), when developing machine learning based models and systems for different consumer and enterprise applications. Then, we will focus on the application of fairness-aware machine learning techniques in practice, by highlighting industry best practices and case studies from different technology companies. Based on our experiences in industry, we will identify open problems and research challenges for the data mining / machine learning community.

ACM Reference Format:
Sarah Bird, Ben Hutchinson, Krishnaram Kenthapadi, Emre Kıcıman, and Margaret Mitchell. 2019. Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320086

1 OUTLINE OF THE TUTORIAL

The tutorial will consist of the following parts: an overview of algorithmic bias / discrimination issues observed in practice over the last few years and the lessons learned (1 hr), fairness notions and techniques from the perspective of different web applications (1 hr), and case studies from different technology companies, along with open problems and research directions (1 hr).

We will cover the following key topics:

- Introduction to algorithmic bias / discrimination
- Industry best practices
- Sources of biases in ML lifecycle
- Algorithmic techniques for fairness in ML
- Fairness methods in practice: Challenges and lessons learned
  - Image and related topics
  - Machine translation
  - Conversational agents
  - Web search, talent search, and other ranking domains
  - Open problems
  - Key takeaways and reflections

1.1 Algorithmic Bias in Machine Learning

The topic of algorithmic bias and discrimination has been studied extensively across disciplines such as law, policy, and computer science (e.g., see [2, 11] and the references therein). An early work considers a computer system to be biased if it systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others [9]. Systematic discrimination combined with an unfair outcome is considered to result in bias.

Many scholars have investigated two different notions of fairness: (1) *individual fairness*, which requires similar treatment for similar people [6], and (2) *group fairness*, which requires similar treatment for the disadvantaged group and the advantaged group [16]. Some studies focus on quantifying the extent of discrimination (e.g., [1, 3, 16]), while others explore fairness-aware algorithms for mitigating bias (e.g., [4–8, 12–15, 17–20]) and limitations / trade-offs in achieving different notions of fairness [5–7, 15].

In the tutorial, we will also discuss sources of data bias at various stages of the data generation, collection, and analysis pipeline, and current best practices for identifying/monitoring and mitigating these sources of bias. We will present examples of algorithmic bias / discrimination issues observed over the last few years, highlighting the pitfalls / lessons learned and strategies to address these issues.

1.2 Fairness Notions

We will discuss different definitions of fairness in machine learning (such as statistical parity, equalized odds, and equal opportunity [13]), and the relative advantages and limitations of these definitions. We will also present a broad taxonomy of web applications, and discuss appropriate fairness notions for each class.

Ranking users for a given "query": Examples include ranking individuals for credit offers and lending, ranking potential candidates in response to a recruiter's search query or job posting, ranking results about people in web or social search, and ranking students for college admissions. Here we are concerned with

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320086

fairness to the users being ranked. Applicable fairness notions include statistical parity [6], equalized odds/opportunity [13], and fairness-aware ranking [4, 19].

Ranking items for a given user: Several common web applications fall into this category, e.g., news article recommendations, job search and recommendations, microtargeted ads, movie recommendations, etc. The items could be ranked based on explicit search query, on user profile, and/or on past user behavior. Here we are concerned with fairness for the user to whom we present the ranked items. For instance, a desirable fairness notion could be to ensure that we present similar items to similar users (individual fairness). In some cases, there may be conflicts to resolve between what the user likes as judged by behavior, and fairness goals. For instance, men and women might have different preferences for specific types of news articles or ads.

Ranking other users for a given user: Examples include YouTube channel recommendations, people recommendations (*People You May Know* of Facebook and LinkedIn, *Who To Follow* of Twitter), and various dating applications. We could be concerned about fairness for both the users that receive recommendations, and the users they are recommended to or not. Also, there may be a fine balance needed between fairness and preferences – for instance, supposing that users are more likely to connect with others of similar age or experience level, one needs to ponder whether to recommend people of different ages to people of different ages.

Other considerations: There may also be fairness considerations that do not belong to the above categories. Consider for example, personalized web applications that send notifications to users (e.g., push notifications in the app, email notification). It may be desirable to ensure that different classes of users are treated similarly with respect to how often and when they are presented with such notifications, especially when the notification corresponds to an opportunity (e.g., a credit offering or a job recommendation). For instance, a model that presents men with an earlier notification of a job opening or more frequent notifications of educational or job opportunities could result in disparate impact to women. Further, different classes of users may experience significantly dissimilar treatments due to the presence of externalities. For example, a provider of educational or employment opportunity may either be not able or need to pay more to reach a class of users (e.g., women) due to the presence of other competing ads.

1.3 Case Studies from Industry

As part of the tutorial, we will also focus on the application of fairness-aware machine learning techniques in practice, by presenting case studies from different technology companies spanning applications such as image/vision, machine translation, conversational agents, and talent search (e.g., representative ranking for LinkedIn talent search [10]). We will discuss AI & ethics initiatives within and across different companies (e.g., Partnership on AI), focusing on technical solutions being adopted to address known issues. Finally, we will discuss open challenges and research directions for the community.

2 BRIEF BIOGRAPHIES OF THE PRESENTERS

Sarah Bird leads strategic projects at the intersection of AI research and products at Microsoft. Her current work focuses on AI Ethics and developing AI responsibly at scale. Ben Hutchinson is a Senior Engineer in Google's Research & Machine Intelligence group, working on AI, fairness, and ethics, in Google's Ethical AI team. His interdisciplinary research includes learning from social sciences to inform the ethical development of AI. Krishnaram Kenthapadi is part of the AI team at LinkedIn, where he leads the fairness, transparency, explainability, and privacy modeling efforts across different LinkedIn applications. He also serves as LinkedIn's representative in Microsoft's AI and Ethics in Engineering and Research (AETHER) Committee. Emre Kıcıman is a Principal Researcher at and co-leads Microsoft Research AI's efforts on AI and its implications for people and society. In addition, his research focuses on causal analysis and data bias in the context of computational social science analyses and decision support systems. Margaret Mitchell is a Senior Research Scientist in Google's Research & Machine Intelligence group, working on AI, multimodality, and ethics, and she currently leads Google's Ethical AI team. Her research involves vision-language, computer vision, and grounded language generation, focusing on how to evolve AI towards positive goals.

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, 2016.
[2] S. Barocas and M. Hardt. Fairness in machine learning. In *NIPS Tutorial*, 2017.
[3] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 2017.
[4] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *ICALP*, 2018.
[5] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 2017.
[6] C. Dwork, M. Hardt, T. Pitassi, and R. Z. Omer Reingold. Fairness through awareness. In *ITCS*, 2012.
[7] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv:1609.07236*, 2016.
[8] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. *arXiv:1802.04422*, 2018.
[9] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3), 1996.
[10] S. C. Geyik and K. Kenthapadi. Building representative talent search at LinkedIn. LinkedIn engineering blog post, Available at https://engineering.linkedin.com/blog/2018/10/building-representative-talent-search-at-linkedin, October 2018.
[11] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *KDD Tutorial on Algorithmic Bias*, 2016.
[12] S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 28(5-6), 2014.
[13] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.
[14] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in reinforcement learning. In *ICML*, 2017.
[15] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.
[16] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, 2008.
[17] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *COLT*, 2017.
[18] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 2017.
[19] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *CIKM*, 2017.
[20] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.

A/B Testing at Scale: Accelerating Software Innovation

Alex Deng
Analysis & Experimentation Team
Microsoft

Somit Gupta
Analysis & Experimentation Team
Microsoft

Pawel Janowski
Analysis & Experimentation Team
Microsoft

Ronny Kohavi
Analysis & Experimentation Team
Microsoft

Jeff Omhover
Analysis & Experimentation Team
Microsoft

Abstract

The Internet and the general digitalization of products and operations provides an unprecedented opportunity to accelerate innovation while applying a rigorous and trustworthy methodology for supporting key product decisions. Developers of connected software, including web sites, applications, and devices, can now evaluate ideas quickly and accurately using controlled experiments, also known as A/B tests. From front-end user-interface changes to backend algorithms, from search engines (e.g., Google, Bing, Yahoo!) to retailers (e.g., Amazon, eBay, Etsy) to social networking services (e.g., Facebook, LinkedIn, Twitter) to travel services (e.g., Expedia, Airbnb, Booking.com) to many startups, online controlled experiments are now utilized to make data-driven decisions at a wide range of companies. The theory of a controlled experiment is simple, but for the practitioner the deployment and evaluation of online controlled experiments at scale (100's of concurrently running experiments) across a variety of web sites, mobile apps, and desktop applications presents many pitfalls and new research challenges. In this tutorial, we will introduce the overall A/B testing methodology, walkthrough use cases using real examples, and then focus on practical and research challenges in scaling experimentation. We will share key lessons learned from scaling experimentation at Microsoft to thousands of experiments per year and outline promising directions for future work.

ACM Reference format:

Alex Deng, Somit Gupta, Pawel Janowski, Ronny Kohavi and Jeff Omhover. 2019. A/B Testing at Scale: Accelerating Software Innovation. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320093

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320093

1 Previous editions

The tutorial in its current or close to its current form has been presented to a large audience before at SIGIR 2017 [1] , KDD 2017, and STRATA 2018 [2]. Some topics covered in the tutorial were discussed in Ronny Kohavi's keynote talk at the KDD 2015 conference [3]. While the keynote covered a range of topics in a brief fashion, the tutorial will go in depth and will also include material from works such as [1] [8] [12] [7] [9] [4]. Requests for a more in-depth tutorial from those who attended the keynote and other conference talks the authors have given over the last two years is one of the key motivations for us to submit this tutorial proposal. Parts of the tutorial are based on the "Introduction to Experimentation" training course Analysis and Experimentation team conducts internally at Microsoft on a monthly basis. Some of the advanced statistical techniques discussed in this tutorial were also presented in a tutorial by Deng et. al at JSM 2015 conference. However, the material has been adapted to not require advanced statistical knowledge as a prerequisite. A tutorial on experimentation on the web was given by Ronny Kohavi at el at KDD 2009 [6]. Since that time, both the theory of online A/B testing and its use in practice have evolved greatly, and the overlap of our current tutorial proposal with that one is not large. Slides and videos from some of the past talks given by authors can be viewed at http://exp-platform.com/talks/

References

[1] P. D. S. G. R. K. P. R. a. L. V. Alex Deng, "A/B Testing at Scale: Accelerating Software Innovation," in *In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*, Shinjuku, Tokyo, Japan , 2017.

[2] R. K. A. D. P. R. Somit Gupta, "A/B Testing at Scale Tutorial," March 2018. [Online]. Available: https://exp-platform.com/2018StrataABtutorial/.

[3] R. Kohavi, "Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years," Conference on Knowledge Discovery and Data Mining (KDD).

[4] A. Fabijan, P. Dmitriev, H. Holmstrom and J. Bosch, "The Evolution of Continuous Experimentation in Software Product Development," in *International Conference on Software Engineering (ICSE)*, 2017.

[5] A. Deng and X. Shi, "Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[6] R. Kohavi, R. Longbotham and J. Quarto-vonTivadar, "Planning, Running, and Analyzing Controlled Experiments on the Web," in *tutorial at Conference on Knowledge Discovery and Data Mining*, 2009.

[7] R. Kohavi, "Pitfalls in Online Controlled Experiments," in *MIT COnference on Digital Experimentation (CODE)*, 2016.

[8] Z. Zhao, M. Chen, D. Matheson and M. Stone, "Online Experimentation Diagnosis and Troubleshooting Beyond AA Validation," in *Conference on Data Science and Advanced Analytics*, 2016.

[9] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker and Y. Xu, "Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[10] A. Deng, Y. Xu, R. Kohavi and T. Walker, "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data," in *Conference on Web Search and Data Mining (WSDM)*, 2013.

[11] P. Dmitriev and X. Wu, "Measuring Metrics," in *Conference on Information and Knowledge Management (CIKM)*, 2016.

[12] W. Machmouchi and G. Buscher, "Principles for the Design of Online A/B Metrics," in *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2016.

[13] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu and N. Pohlmann, "Online Controlled Experiments at Large Scale," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[14] A. Deng, "Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments," in *World Wide Web Conference (WWW)*, 2015.

[15] A. Deng, J. Lu and S. Chen, "Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing," in *Conference on Data Science and Advanced Analytics*, 2016.

[16] A. Deng, P. Zhang, S. Chen, D. Kim and J. Lu, "Concise Summarization of Heterogeneous Treatment Effect Using Total Variation Regularized Regression," in *In submission*, 2017.

[17] R. Kohavi, "SPEED MATTERS: A SLOWDOWN EXPERIMENT," [Online]. Available: https://1drv.ms/w/s!AuRxCGEOCRKGkbtoabZo0OcKoo6xhw.

Economic Theories of Distributive Justice for Fair Machine Learning

Krishna P. Gummadi
MPI-SWS, Saarbrücken
gummadi@mpi-sws.org

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

ABSTRACT

Machine Learning is increasingly employed to make consequential decisions for humans. In response to the ethical issues that may ensue, an active area of research in ML has been dedicated to the study of algorithmic unfairness. This tutorial introduces fair-ML to the web conference community and offers a new perspective on it through the lens of the long-established economic theories of distributive justice. Based on our past and ongoing research, we argue that economic theories of equality of opportunity, inequality measurement, and social choice have a lot to offer—in terms of tools and insights—to data scientists and practitioners interested in understanding the ethical implications of their work. We overview these theories and discuss their connections to fair-ML.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; • **Applied computing** → **Economics**.

KEYWORDS

social choice theory; equality of opportunity; inequality

ACM Reference Format:
Krishna P. Gummadi and Hoda Heidari. 2019. Economic Theories of Distributive Justice for Fair Machine Learning. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320101

1 INTRODUCTION

Automated data-driven decision-making tools are increasingly employed to make consequential decisions for human subjects— examples include employment [19], credit lending [20], criminal justice [1], policing [23], and medicine [14]. Decisions made in this fashion can have a long-lasting impact on society and may affect certain individuals or groups negatively [25]. This realization has recently spawned a new active area of research into quantifying and guaranteeing fairness for machine learning [4, 12, 17].

As concerns over algorithmic unfairness and discrimination continue to grow in magnitude and depth, it is timely—and critical—for data scientists and practitioners to be armed with a toolbox of models and mechanisms to quantify and tackle algorithmic unfairness in their application domains. This tutorial offers a new perspective on fair ML through the lens of the long-established *economic* theories of distributive justice. The extensive economic literature on *equality of opportunity*, *measurement of inequality*, *social choice theory*, and *fair division* has a lot to offer—in terms of models and tools—to data scientists and practitioners, interested in understanding the ethical implications of their work.

The main objectives of this tutorial are:

- to overview the growing line of work on fairness for ML;
- to survey the economic theories of distributive justice;
- to put Fair ML into economic perspective and terminology;
- to cast the well-established economics of distributive justice as a blueprint to guide and inform future research into fairness for Machine Learning.

We begin with an overview of the fair ML literature. We introduce existing notions of algorithmic (un)fairness and summarize some of the major themes and findings in Fair ML. We then cast these notions as special cases of economic models of Equality of Opportunity (EOP). Through this lens, we offer a better understanding of the moral assumptions underlying technical definitions of fairness. Second, we discuss the conception of unfairness as inequality. We overview the axiomatic characterization of measures of (income) inequality and present them as a unifying framework for quantifying both individual- and group-level unfairness. Third, we discuss the "leveling down objection" to equality. We propose the use of cardinal social welfare functions to address this issue and as an efficient method for bounding inequality. Last but not least, we discuss how differing (group) preferences can justify unequal outcomes, drawing on the concepts of envy and equity from fair allocation.

2 SCOPE

This tutorial discusses several seminal papers from economics (e.g., [2, 3, 13, 18]) and connects them to the growing body of work on fairness for ML. We overview of the recent research on algorithmic fairness. We introduce group-level notions of fairness—which require that given a classifier, a certain fairness metric is equal across all protected groups (see e.g. [17, 27, 28])—-as well as individual-level notions of fairness [9]—which requires that two individuals who are similar with respect to the task at hand receive similar classification outcomes.

Fair ML as Equality of Opportunity: Next, we offer a new moral framework for understanding these notions by mapping them to economic models of Equality of opportunity (EOP) [16]. EOP is a widely supported ideal of fairness, and it has been extensively studied in political philosophy and economics [22]. We show that through this conceptual mapping, many existing definitions of algorithmic fairness, such as predictive value parity and equality odds, can be interpreted as special cases of EOP. This approach

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320101

allows us to explicitly spell out the moral assumptions underlying each notion of fairness. Moreover, it confers a moral meaning to the recent fairness impossibility results.

(Un)fairness as (In)equality: The understanding of fairness as some form of *equality* is the fundamental basis for many theories of justice. Given the distribution of individuals' *well-being* or *benefits*, measures of inequality capture the degree to which the distribution is unequal and allow for a direct comparison of inequality between various distributions. We start by noting that parity-based definitions of algorithmic fairness, such as statistical parity [5], disparate impact [10], and equality of opportunity [12], can be thought of as seeking to minimize some form of inequality—for different notions of *benefit* or *well-being*. This observation motivates our interest in inequality measurement. We overview the axiomatic characterization of inequality indices, using Gini, Theil, and Generalized entropy indices as examples [6, 8]. We discuss these axioms—such as *population-invariance* and *progressive transfer* principle—in the context of fair ML. We argue that inequality indices can be utilized to extend existing definitions of algorithmic fairness to multiple groups and settings beyond binary classification. Furthermore, a particular structural property of these indices, called *additive decomposability*, allows us to interpolate between individual and group-level (un)fairness and observe the tradeoffs between the two [24]. The main challenges we see in utilizing inequality indices as measures of algorithmic unfairness are (a) defining the right notion of well-being/benefit to equalize across groups/individuals, and (b) finding efficient and precise mechanisms for bounding inequality.

Next, we overview two economic theories of distributive justice that depart from the idea of fairness as equality: social choice theory and fair division. Social choice theory is concerned with aggregating individual preferences to pick a *just* collective outcome, where justice is defined through a set of axioms. Fair division is concerned with the division of a limited resource (e.g., a cake) among individuals with heterogeneous valuations/preferences.

Social Choice Theory: When individual utilities are interpersonally comparable, a cardinal social welfare function can be designed to choose a collective outcome [13]. Social welfare functions can be interpreted as measures of distributive justice behind a *veil of ignorance* [21]. This interpretation motivates our interest in employing them to measure algorithmic fairness. We go over the axiomatic characterization of cardinal social welfare functions, discussing such axioms as *symmetry* and *independence of unconcerned agents* in the context of fair ML. We next note that according to the Debreu-Gorman theorem [7, 11], the family of social welfare functions satisfying the above axioms is strikingly small. Furthermore and unlike measures of inequality, this class of social welfare functions enjoys a convex formulation. Therefore, it can be readily integrated into any convex loss minimization pipeline [15]. Last but not least, we briefly discuss the connections between welfare and inequality aversion through a welfare-based interpretation of the Atkinson's measure of inequality [3] and show that guaranteeing high social welfare usually leads to low inequality in practice.

Fair Division: Fair division accounts for the fact that different entities may have different preferences for different outcomes. We formally introduce the notions of *envy* and *equity*. An allocation is *envy-free* if no individual prefers the allocation of another to their own [18, 26]. We discuss one recent adaptation of no-envy to capture algorithmic fairness at the group level [29].

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Propublica* (2016).
[2] Kenneth J. Arrow. 2012. *Social choice and individual values*. Vol. 12. Yale university press.
[3] Anthony B. Atkinson. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 3 (1970), 244–263.
[4] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* (2016).
[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDM workshops*. 13–18.
[6] Frank A. Cowell. 2000. Measurement of inequality. *Handbook of Income Distribution* 1 (2000), 87–166.
[7] Gerard Debreu. 1959. *Topological methods in cardinal utility theory*. Technical Report. Cowles Foundation for Research in Economics, Yale University.
[8] Bhaskar Dutta. 2002. Inequality, poverty and welfare. *Handbook of social choice and welfare* 1 (2002), 597–633.
[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*. ACM, 214–226.
[10] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of KDD*. ACM, 259–268.
[11] William M. Gorman. 1968. The structure of utility functions. *The Review of Economic Studies* 35, 4 (1968), 367–390.
[12] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of NIPS*. 3315–3323.
[13] John C. Harsanyi. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy* 63, 4 (1955), 309–321.
[14] Robert David Hart. 2017. If you're not a white male, artificial intelligence's use in healthcare could be dangerous. *Quartz* (July 2017).
[15] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Proceedings of NeurIPS*.
[16] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)*.
[17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *In proceedings of the 8th Innovations in Theoretical Computer Science Conference*.
[18] Serge-Christophe Kolm. 2002. *Justice and equity*. MIT Press.
[19] Claire Miller. 2015. Can an Algorithm Hire Better than a Human? *The New York Times* (June 25 2015).
[20] Kevin Petrasic, Benjamin Saul, James Greig, and Matthew Bornfreund. 2017. Algorithms and bias: What lenders need to know. *White & Case* (2017).
[21] John Rawls. 2009. *A theory of justice*. Harvard university press.
[22] John E. Roemer and Alain Trannoy. 2015. Equality of opportunity. In *Handbook of income distribution*. Vol. 2. Elsevier, 217–300.
[23] Cynthia Rudin. 2013. Predictive Policing Using Machine Learning to Detect Patterns of Crime. *Wired Magazine* (August 2013).
[24] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual and Group Unfairness via Inequality Indices. In *Proceedings of KDD*.
[25] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.
[26] Hal R. Varian. 1974. Equity, envy, and efficiency. *Journal of economic theory* 9, 1 (1974), 63–91.
[27] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of WWW*. 1171–1180.
[28] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.
[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Proceedings of NIPS*. 228–238.

Tutorial on Online User Engagement: Metrics and Optimization

Extended Abstract

Liangjie Hong
Etsy Inc.
lhong@etsy.com

Mounia Lalmas
Spotify
mounia@acm.org

ABSTRACT

User engagement plays a central role in companies operating online services, such as search engines, news portals, e-commerce sites, entertainment services, and social networks. A main challenge is to leverage collected knowledge about the daily online behavior of millions of users to understand what engage them short-term and more importantly long-term. Two critical steps of improving user engagement are metrics and their optimization. The most common way that engagement is measured is through various online metrics, acting as proxy measures of user engagement. This tutorial will review these metrics, their advantages and drawbacks, and their appropriateness to various types of online services. Once metrics are defined, how to optimize them will become the key issue. We will survey methodologies including machine learning models and experimental designs that are utilized to optimize these metrics via direct or indirect ways. As case studies, we will focus on four types of services, news, search, entertainment, and e-commerce. We will end with lessons learned and a discussion on the most promising research directions.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Online shopping**; • **General and reference** → *Evaluation*; *Experimentation*; • **Mathematics of computing** → *Hypothesis testing and confidence interval computation.*

KEYWORDS

User Engagement, Evaluation, Online Experiments, Counterfactual Learning

ACM Reference Format:
Liangjie Hong and Mounia Lalmas. 2019. Tutorial on Online User Engagement: Metrics and Optimization: Extended Abstract. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3320087

1 MOTIVATION

In the online world, user engagement refers to the quality of the user experience that emphasizes the positive aspect of the interaction with an online service and, in particular, the phenomena associated with wanting to use that service longer and more frequently [1].

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320087

User engagement is a key concept in the design of online services, as successful applications are not just used, but are engaged with. Users invest time, attention, and emotion with services they use [11]. How to measure user engagement and being able to drive it becomes a key element for evaluating the success of any online services, from news to e-commerce sites, as it informs on the understanding of user needs and expectations, system design and functionality.

User engagement is a multifaceted, complex phenomenon, and as such has given rise to a number of approaches for its measurement [11]. Common ways of measuring user engagement include: self-reporting, e.g., questionnaires; observational methods, e.g., desktop actions; and online analytics, which employ online behavior metrics that assess users' depth of engagement. These approaches represent various trade-offs between the scale of data analyzed and the depth of understanding. In this tutorial, we focus on the later, large-scale online analytics. Within this approach, the most common way that engagement is measured is through various proxy measures of user engagement. Standard metrics include number of page views, number of unique users, dwell time, bounce rate, click-through rate, and return rate. We will review these metrics, and discuss what they measure, their advantages and drawbacks. We will provide extensive details on the appropriateness of these metrics to various types of online services.

Once metrics are identified and measured, it becomes a challenge to derive a mechanism to optimize them. From a machine learning perspective, two types of approaches are exploited to drive metrics, directly optimizing a metric when it can be formalized easily in a mathematical form, or indirectly through some surrogate objective functions. From an experimental design perspective, certain metrics can be easily observed and improved during a period of time of an online A/B experiment in terms of several weeks while the change of other metrics, especially long term ones, can be difficult or even never observed and improved directly through experiments. In this tutorial, we will discuss these aspects and explore new research directions.

Throughout the tutorial, we will use applications in areas we have extensive experiences, news, search, entertainment, and e-commerce to demonstrate how online user engagement can be measured and optimized.

2 BIOGRAPHY

Liangjie Hong is Director of Engineering, Data Science and Machine Learning at Etsy Inc., managing a group of data scientists to deliver cutting-edge scientific solutions for: Search and Discovery, Personalization & Recommendation and Computational Advertising. Previously, he was Senior Manager of Research at Yahoo Research, leading science efforts for Personalization and Search Sciences. Prior to Yahoo Research, he obtained his PhD in Computer

Science from Lehigh University in 2013. Liangjie has given numerous technical talks at academic conferences as well as industrial meetings. He also co-founded User Engagement Optimization Workshop,[1] which has been held in conjunction with CIKM 2013 and KDD 2014. Liangjie has extensively published papers in optimizing user engagement in the context of personalization and recommendation systems, including designing metrics, developing machine learning methodologies as well as novel models (e.g., [3, 15, 18, 19]). He has served as program committee members on all major applied machine learning and data mining conferences including KDD, WSDM, WWW, SIGIR, EMNLP and ICML.

Mounia Lalmas is a Director of Research at Spotify, and the Head of Tech Research in User Engagement, where she leads an interdisciplinary team of research scientists in Boston, London, New York and San Francisco, working on personalization and discovery. Before that, she was a Director of Research at Yahoo, where she led a team of researchers working on advertising quality for Gemini, Yahoo native advertising platform. She also worked with various teams at Yahoo on topics related to user engagement in the context of news, search, and user generated content. She also hold an Honorary Professorship at University College London. Prior to this, she held a Microsoft Research/RAEng Research Chair at the School of Computing Science, University of Glasgow. Before that, she was Professor of Information Retrieval at the Department of Computer Science at Queen Mary, University of London. Since 2011, her work has focused on studying user engagement [12] in areas such as native advertising [2], digital media [8], user generated content [13], search [4] and lately in music [14]. She has given numerous talks and tutorials on these and related topics. She is regularly a senior programme committee member at conferences such as WSDM, WWW and SIGIR. She was co-programme chair for SIGIR 2015 and WWW 2018. She is also the co-author of a book written as the outcome of her WWW 2013 tutorial [11].

3 DETAILED OUTLINE

The two presenters are experts in user engagement and have published widely on this topic (see the References section). They have worked on products, focusing on understanding users to devise metrics to assess their depth of engagement, and influence the design of algorithms aiming at providing more engaging experience. The tutorial will be based around their work and other relevant works, focusing on large-scale measurement and optimization of online services. The tutorial will be organized around the following lines:

(1) We will start with an introduction to the tutorial, presenters and the scope of the tutorial. We motivate the tutorial by emphasizing that user engagement optimization is the key to successful modern data-driven products.

(2) We will provide an overview of a wide range of user engagement metrics, when to use them and why, their pros and cons. We will provide a taxonomy of metrics, for example, whether they reflect the during-the-moment interaction (e.g. dwell time) or popularity (e.g. number of sessions). We will also provide a classification of these metrics in terms of the types of services for which they are the most appropriate (e.g. dwell time on news portal). Although the focus of this tutorial is on large-scale online analytics, we will briefly discuss the usage of mixed methods, as a mean to inform the choice and design of online engagement metrics through a better understanding of user expectations and experiences with services e.g. [5].

(3) We will discuss the importance in understanding the relationship between short-term and long-term metrics, as the means to ensure we are measuring and optimizing for the right thing. Long-term metrics are usually what we really want to be successfully measuring, like how frequently users come back to the service, how many monthly active users a service would have and how much incremental revenue they could bring in for the service. However, they are hard to measure and difficult to directly optimize for. On the other hand, we may able to build machine learning models and algorithms that optimize for short-term engagement metrics (e.g. click-through rates or dwell time), but we need to be sure that by doing so actually leads to optimizing long-term engagement metrics.

(4) After generic discussions on user engagement metrics, we focus on user engagement optimization on four important online applications: news portals, entertainment streaming services (music, video), search engines, and e-commerce sites. These four cover a large proportions of online services, and differ in the type of engagement they seek [12]; e.g. a successful experience with a search engine is one when the user enters his or her query, get the results fast, and is happy with them and as such is more likely to return to that search engine when he or she has another information need; whereas for a news portal, success means users spending time on the portal, and consuming many articles, every morning or during lunch time. For a music streaming service, the time spent successfully finding what to listen to now or later and listening time are important indicators of lean-in and lean-back engagement modes with these types of services. Finally, a satisfied e-commerce user experience may contain both the ability to help users finding relevant products and completing their transactions, and to lead users to discover new interesting things they would not have considered in the first place. The presenters have also worked on these types of online services.

(5) We then discuss how to optimize user engagement metrics from both machine learning and experimental design perspectives. In particular, we are going to show cases where metrics can be directly optimized like [16] and surrogate objective functions need to be found such as applications in news recommendation [18] and mobile search applications [6]. We will also show how to design and carry out online experiments (e.g. so called A/B testing) and explore frontier directions where sequential decision making and experimental design can be exploited to optimize metrics online like [17]. When online experiments are not always

[1] http://www.ueo-workshop.com/

possible, an emerging direction called counterfactual learning allows us perform offline A/B testing in large scale. We will briefly review its applications in search [7].

(6) We will end with some concluding remarks and future directions of research. We will highlight what are the main themes we believe we should be focusing on next, for example: metrics specifically tailored to engagement on mobile apps, as well as accounting for cross-device engagement and voice interactions; counterfactual analysis for offline online evaluation and the relationship between online and offline evaluation, which allows prioritising A/B tests; and the correct interpretability of online metrics (e.g. short dwell time is not necessarily bad whereas high click-through rate can be bad in terms of engagement). Finally, we will briefly discuss online engagement and its measurement with respect to conversational tools, as many services are promoting the usage of chat bots, and what engagement means in this context is not clear.

4 PREVIOUS EDITION

A former edition of the tutorial entitled "Tutorial on Metrics of User Engagement" has been held at WSDM 2018 [9], which was the third most poupular tutorial at the conference. The slides as of today (submission deadline date), which are available online have[2], have more than 4500 views. In that tutorial, as applications, we focused on search and e-commerce. We propose to add news and entertainment (e.g. music, video) as they bring a different perspective to online user engagement, being about consuming content, as opposed to fulfilling a need (information or shopping). New to this edition is we intend to spend time on reflections and future directions. We feel that with our extensive expertise, we are well positioned to reflect on where we are, what are the important lessons learned, and where should be the main focus next be, both in terms of improving what is being done now and new directions.

5 DURATION AND AUDIENCE

We propose the duration of the tutorial as half-day with approximated 4 hours. The tutorial targets audience of academic researchers and industrial practitioners who are interested in understanding all aspects of user engagement including metric design, measurement and optimization.

6 MATERIALS SUPPLIED TO ATTENDEES

We will provide detailed slides to attendees. Examples of the types of slides can be seen on the above mentioned tutorials "Metrics of user engagement – Applications to Search & E-commerce" [9] (first edition of this tutorial) and "Measuring User Engagement" [10]. We will also provide a detailed annotated bibliography, which will be new with this second edition of the proposed tutorial.

REFERENCES

[1] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. 2011. Towards a science of user engagement. In *WSDM'11*. ACM.

[2] https://www.slideshare.net/mounialalmas/tutorial-on-metrics-of-user-engagement-applications-to-search-e-commerce

[2] Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. 2016. Improving Post-Click User Engagement on Native Ads via Survival Analysis. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. 761–770.

[3] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On Sampling Strategies for Neural Network-based Collaborative Filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 767–776.

[4] Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*. 173–182.

[5] Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas, Ben Carterette, and Fernando Diaz. 2018. Mixed methods for evaluating user satisfaction. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. 541–542.

[6] Liangjie Hong, Yue Shi, and Suju Rajan. 2016. Learning Optimal Card Ranking from Query Reformulation. *CoRR* abs/1606.06816 (2016).

[7] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, New York, NY, USA, 781–789.

[8] Dmitry Lagun and Mounia Lalmas. 2016. Understanding User Attention and Engagement in Online News Reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*. 113–122.

[9] Mounia Lalmas and Liangjie Hong. 2018. Tutorial on Metrics of User Engagement: Applications to News, Search and E-Commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 781–782. https://www.slideshare.net/mounialalmas/tutorial-on-metrics-of-user-engagement-applications-to-search-e-commerce

[10] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. 2013. Measuring User Engagement. https://www.slideshare.net/mounialalmas/measuring-userengagement Tutorial presented at World Wide Web Conference, Rio de Janeiro, Brazil.

[11] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. 2014. *Measuring User Engagement*. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers.

[12] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of User Engagement. In *User Modeling, Adaptation, and Personalization - 20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012. Proceedings*. 164–175.

[13] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. 2014. Reader preferences and behavior on Wikipedia. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*. 88–97.

[14] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. 2243–2251.

[15] Yue Ning, Yue Shi, Liangjie Hong, Huzefa Rangwala, and Naren Ramakrishnan. 2017. A Gradient-based Adaptive Learning Framework for Efficient Personal Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 23–31.

[16] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning Clicks into Purchases: Revenue Optimization for Product Search in E-Commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 365–374.

[17] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. Returning is Believing: Optimizing Long-term User Engagement in Recommender Systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 1927–1936.

[18] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond Clicks: Dwell Time for Personalization. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 113–120.

[19] Qian Zhao, Yue Shi, and Liangjie Hong. 2017. GB-CENT: Gradient Boosted Categorical Embedding and Numerical Trees. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1311–1319.

Modeling and Mining Feature-Rich Networks

Rushed Kanawati
LIPN, University Paris 13
Paris, France
rushed.kanawati@lipn.fr

Martin Atzmueller
Tilburg University, Department of
Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
m.atzmuller@uvt.nl

ABSTRACT

In the field of web mining and web science, as well as data science and data mining there has been a lot of interest in the analysis of (social) networks. With the growing complexity of heterogeneous data, feature-rich networks have emerged as a powerful modeling approach: They capture data and knowledge at different scales from multiple heterogeneous data sources, and allow the mining and analysis from different perspectives. The challenge is to devise novel algorithms and tools for the analysis of such networks.

This tutorial provides a unified perspective on feature-rich networks, focusing on different modeling approaches, in particular multiplex and attributed networks. It outlines important principles, methods, tools and future research directions in this emerging field.

CCS CONCEPTS

• **Theory of computation** → **Graph algorithms analysis**; • **Human-centered computing** → *Social networks*; *Social media*; • **Information systems** → *World Wide Web*.

KEYWORDS

social media, social network analysis, social interaction networks, feature-rich networks, mining social networks

ACM Reference Format:
Rushed Kanawati and Martin Atzmueller. 2019. Modeling and Mining Feature-Rich Networks. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320098

1 INTRODUCTION

In the last years, networks have proven to be a useful tool for modeling structural complexity of a variety of complex systems in different domains including sociology, biology, ethology and computer science. In contrast to simple homogeneous static networks, networks in real world settings are often dynamic, heterogeneous, and both nodes and links can be described by a set of attributes and/or multiple relations. Thus, the mining of such *feature-rich* networks [20] is gaining increasing interest; such networks include, in particular, *node-attributed* and/or *link-attributed* networks, including multiplex networks. This tutorial provides a survey on the latest algorithmic advances for modeling and mining feature-rich networks, including both node and link attributed networks.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320098

We structure the materials as follows: First, we start with a brief introduction defining the field of feature-rich networks. Then, two main sections are presented : the first on modeling and mining *multiplex networks* and the second on mining *attributed networks*. Then, we present applications of the presented network mining methods, and brief overview on available analysis tools. We conclude with a summary and interesting research directions.

2 MODELING AND MINING MULTIPLEX NETWORKS

The concept of multiplex networks has been introduced with the goal to provide an expressive model for modeling real-world complex networks [11, 23]. A multiplex network is roughly defined as a multi-layer graph where each layer contains the same set of nodes but interconnected by different types of links. Nodes across different layers can also be coupled by special *coupling links*. Different types of coupling links can be defined allowing to model various types of networks including; multi-relational networks, dynamic networks or heterogeneous networks. Analysis of multiplex networks requires redefining most of the basic concepts and metrics usually used for plain complex networks including: nodal degree and neighborhood, paths, as well as different centrality functions. [9, 11, 27]. It also requires providing new algorithms to handle basic complex networks analysis tasks such as community detection [18] and link prediction [36]. We cover all these issues with a special focus on the central problem of community detection.

3 MODELING AND MINING ATTRIBUTED NETWORKS

Attributed networks, where nodes and/or links are labeled with additional information, allow further dimensions for network analysis and mining, e. g., [34, 43, 44]. Then, these enable novel methods and approaches for complex network analysis that directly work on the feature-rich attribute structure. These include approaches, for example, that specifically exploit the descriptive information of the labels assigned to nodes and/or edges of the graph. Nodes denoting actors, for example, can be labeled with properties of the latter for compositional subgroup analysis, which can be implemented using subgroup discovery [1, 2]. Furthermore, explicit descriptions for model building and explication can be provided, e. g.,using community detection methods e. g., [25, 33]. This also extends to link prediction [26]. Here, we provide an overview on recent algorithmic advances for modeling and mining attributed networks. We start with an introduction, providing basic definitions and introductory examples. After that, we tackle two specific methods: community detection, e. g., [3, 5, 6, 8, 10, 12, 17, 32, 35, 41, 42] and link prediction, e. g., [16, 21] on attributed networks.

4 APPLICATIONS AND TOOLS

There are different types of applications of the presented approaches for multiplex and attributed network mining. One important area is recommender systems, for which we show two applications: mining multiplex networks and attributed networks for enhancing tag recommendation in folksonomies [19], and mining multiplex networks for enhancing social interactions at social events [38, 39], respectively. Another original application of multiplex network mining includes an ensemble approach for cluster ensemble selection [22]. Furthermore, a summary of the available tools for multiplex and attributed network mining will be made. This includes a brief review of the Muna Library [14] and the VIKAMINE toolkit [4].

5 CONCLUSIONS

Network science is now mature enough to move towards analyzing and mining networks with much richer information than plain complex networks, modeled as *feature-rich complex networks*. Different extensions and future research directions can be identified: These include hybrid combinations of attributed, multiplex, and dynamic networks, e. g., [38], adaptations of Bayesian approaches for modeling and mining feature-rich networks [7, 13], extending and adapting network embedding methods and deep learning approaches for feature-rich networks, c. f., [15, 24, 37, 40] for some initial directions, and also including semantics into feature-rich network modeling and mining e. g., [28–31].

REFERENCES

[1] Martin Atzmueller. 2015. Subgroup Discovery. *WIREs DMKD* 5, 1 (2015), 35–49.

[2] Martin Atzmueller. 2018. Compositional Subgroup Discovery on Attributed Social Interaction Networks. In *Proc. DS*. Springer, Heidelberg, Germany.

[3] Martin Atzmueller, Stephan Doerfel, and Folke Mitzlaff. 2016. Description-Oriented Community Detection using Exhaustive Subgroup Discovery. *Information Sciences* 329 (2016), 965–984.

[4] Martin Atzmueller and Florian Lemmerich. 2012. VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In *Proc. ECML/PKDD (LNAI)*, Vol. 7524. Springer, Heidelberg, Germany.

[5] Martin Atzmueller and Florian Lemmerich. 2013. Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. *International Journal of Web Science* 2, 1/2 (2013), 80–112.

[6] Martin Atzmueller and Folke Mitzlaff. 2011. Efficient Descriptive Community Mining. In *Proc. FLAIRS*. AAAI, 459 – 464.

[7] Martin Atzmueller, Andreas Schmidt, Benjamin Kloepper, and David Arnu. 2017. HypGraphs: An Approach for Analysis and Assessment of Graph-Based and Sequential Hypotheses. In *New Frontiers in Mining Complex Patterns (LNAI)*, Vol. 10312. Springer, Heidelberg, Germany.

[8] Martin Atzmueller, Henry Soldano, Guillaume Santini, and Dominique Bouthinon. 2018. MinerLSD: Efficient Local Pattern Mining on Attributed Graphs. In *Proc. 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*.

[9] Federico Battiston, Vincenzo Nicosia, and Vito Latora. 2013. Metrics for the analysis of multiplex networks. *CoRR* abs/1308.3182 (2013).

[10] Cécile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenkova. 2015. Clustering Attributed Graphs: Models, Measures and Methods. *Network Science* 3, 3 (2015), 408–444.

[11] Peter Brodka and Przemysaw Kazienko. 2014. *Encyclopedia of Social Network Analysis and Mining*. Springer, Chapter Multi-layered social networks.

[12] David Combe, Christine Largeron, Mathias Géry, and Elöd Egyed-Zsigmond. 2015. I-Louvain: An Attributed Graph Clustering Method. In *Proc. IDA*. 181–192.

[13] Lisette Espín-Noboa, Florian Lemmerich, Markus Strohmaier, and Philipp Singer. 2017. JANUS: A Hypothesis-Driven Bayesian Approach for Understanding Edge Formation in Attributed Multigraphs. *Applied Network Science* 2, 1 (2017), 16.

[14] Issam Falih and Rushed Kanawati. 2015. Muna: A Multiplex network analysis Library. In *The 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Paris, 757–760.

[15] Hongchang Gao and Heng Huang. 2018. Deep Attributed Network Embedding. In *Proc. IJCAI*. 3364–3370.

[16] Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine (Runting) Shi, and Dawn Song. 2014. Joint Link Prediction and Attribute Inference Using a Social-Attribute Network. *ACM Trans. Intell. Syst. Technol.* 5, 2 (2014), 27:1–27:20.

[17] Stephan Günnemann, Ines Färber, Brigitte Boden, and Thomas Seidl. 2013. GAMer: A Synthesis of Subspace Clustering and Dense Subgraph Mining. In *Knowledge and Information Systems*. Springer.

[18] Manel Hmimida and Rushed Kanawati. 2015. Community detection in multiplex networks: a seed-centric approach. *Networks and Heterogeneous Media* 10, 1 (March 2015), 71–85.

[19] Manel Hmimida and Rushed Kanawati. 2016. A graph-coarsening approach for tag recommendation. In *25th international World Wide Web Conference*. 43–44.

[20] Roberto Interdonato, Martin Atzmueller, Sabrina Gaito, Rushed Kanawati, Christine Largeron, and Alessandra Sala. 2019. Feature-rich networks: going beyond complex network topologies. *Applied Network Science* 4, 1 (2019), 4:1–4:13. https://doi.org/10.1007/s41109-019-0111-x

[21] Maosheng Jiang, Yonxiang Chen, and Ling Chen. 2015. Link Prediction in Networks with Node Attributes by Similarity Propagation. *arXiv preprint arXiv:1502.04380* (2015).

[22] R. Kanawati. 2019. Cluster-ensemble selection: an ensemble approach. In *Proc. SCML2019*. Wuhan, China.

[23] Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. 2014. Multilayer Networks. *Journal of complex networks* 2 (2014), 203–271. http://arxiv.org/abs/1309.7233

[24] Jihwan Lee and Sunil Prabhakar. 2018. A3embed: Attribute Association Aware Network Embedding. In *Proc. WWW (Companion)*. IW3C2, 1243–1251.

[25] Jure Leskovec, Kevin J. Lang, and Michael Mahoney. 2010. Empirical Comparison of Algorithms for Network Community Detection. In *WWW*. ACM, 631–640.

[26] David Liben-Nowell and Jon M. Kleinberg. 2007. The link-prediction problem for social networks. *JASIST* 58, 7 (2007), 1019–1031.

[27] Matteo Magnani and Moreno Marzolla. 2014. Path-Based and Whole-Network Measures. In *Encyclopedia of Social Network Analysis and Mining*. 1256–1269. https://doi.org/10.1007/978-1-4614-6170-8_241

[28] Spyroula Masiala and Martin Atzmueller. 2018. First Perspectives on Explanation in Complex Network Analysis. In *Proc. BNAIC*. Jheronimus Academy of Data Science, Den Bosch, The Netherlands.

[29] Spyroula Masiala and Martin Atzmueller. 2018. Towards Explainable Complex Network Analysis". In *Proc. Dutch-Belgian Database Day*. Hasselt University, Belgium.

[30] Folke Mitzlaff, Martin Atzmueller, Andreas Hotho, and Gerd Stumme. 2014. The Social Distributional Hypothesis. *Journal of Social Network Analysis and Mining* 4, 216 (2014).

[31] Folke Mitzlaff, Martin Atzmueller, Gerd Stumme, and Andreas Hotho. 2013. Semantics of User Interaction in Social Media. In *Complex Networks*. Springer, Heidelberg, Germany.

[32] Flavia Moser, Recep Colak, Arash Rafiey, and Martin Ester. 2009. Mining Cohesive Patterns from Graphs with Feature Vectors. In *SDM*, Vol. 9. SIAM, 593–604.

[33] M. E. J. Newman. 2006. Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences* 103, 23 (2006), 8577–8582.

[34] Joseph J Pfeiffer III, Sebastian Moreno, Timothy La Fond, Jennifer Neville, and Brian Gallagher. 2014. Attributed Graph Models: Modeling Network Structure with Correlated Attributes. In *Proceedings of the 23rd international conference on World wide web*. ACM, 831–842.

[35] S. Pool, F. Bonchi, and M. van Leeuwen. 2014. Description-driven Community Detection. *TIST* 5, 2 (2014).

[36] Manisha Pujari and Rushed Kanawati. 2017. *Interactions in complex systems*. Cambridge Scholar Publishing, Chapter Link prediction in Large-scale multiplex networks.

[37] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *European Semantic Web Conference*. Springer, 593–607.

[38] Christoph Scholz, Martin Atzmueller, Alain Barrat, Ciro Cattuto, and Gerd Stumme. 2013. New Insights and Methods For Predicting Face-To-Face Contacts. In *Proc. AAAI ICWSM*. AAAI Press, Palo Alto, CA, USA.

[39] Christoph Scholz, Martin Atzmueller, and Gerd Stumme. 2012. On the Predictability of Human Contacts: Influence Factors and the Strength of Stronger Ties. In *Proc. SocialCom*. IEEE Computer Society, Boston, MA, USA.

[40] Nasrullah Sheikh, Zekarias Kefato, and Alberto Montresor. 2018. GAT2VEC: Representation Learning for Attributed Graphs. *Computing* (2018), 1–23.

[41] Arlei Silva, Wagner Meira Jr, and Mohammed J Zaki. 2012. Mining Attribute-Structure Correlated Patterns in Large Attributed Graphs. *Proc. VLDB Endowment* 5, 5 (2012), 466–477.

[42] Jaewon Yang, Julian McAuley, and Jure Leskovec. 2013. Community Detection in Networks with Node Attributes. *2013 IEEE ICDM* (2013), 1151–1156. arXiv:1401.7267

[43] Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. 2010. LINKREC: A Unified Framework for Link Recommendation with User Attributes and Graph Structure. In *Proc. WWW*. 1211–1212.

[44] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2009. Graph Clustering Based on Structural/Attribute Similarities. *Proc. VLDB Endow.* 2, 1 (2009), 718–729.

Privacy-preserving Data Mining in Industry

Krishnaram Kenthapadi
LinkedIn
kkenthapadi@linkedin.com

Ilya Mironov
Google
mironov@google.com

Abhradeep Guha Thakurta
UC Santa Cruz
aguhatha@ucsc.edu

ABSTRACT

Preserving privacy of users is a key requirement of web-scale data mining applications and systems such as web search, recommender systems, crowdsourced platforms, and analytics applications, and has witnessed a renewed focus in light of recent data breaches and new regulations such as GDPR. In this tutorial, we will first present an overview of privacy breaches over the last two decades and the lessons learned, key regulations and laws, and evolution of privacy techniques leading to differential privacy definition / techniques. Then, we will focus on the application of privacy-preserving data mining techniques in practice, by presenting case studies such as Apple's differential privacy deployment for iOS / macOS, Google's RAPPOR, LinkedIn Salary, and Microsoft's differential privacy deployment for collecting Windows telemetry. We will conclude with open problems and challenges for the data mining / machine learning community, based on our experiences in industry.

ACM Reference Format:
Krishnaram Kenthapadi, Ilya Mironov, and Abhradeep Guha Thakurta. 2019. Privacy-preserving Data Mining in Industry. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3320085

1 OUTLINE OF THE TUTORIAL

1.1 Privacy Breaches and Lessons Learned

We will present the key privacy breaches over the last two decades, highlighting the attacker's advantage and the lessons learned in each case. The examples to be presented include the following.

- Sweeney's de-anonymization of MA governor's health records [25] (Lesson learned: The attacker can make use of auxiliary information (which can sometimes come from unexpected sources).)
- AOL search log release [2] (Lesson learned: Significant damage can be inflicted even if the attacker is able to succeed on a small fraction of the inputs.)
- De-anonymizing Netflix data [20] and de-anonymizing web browsing data with social networks [24] (Lesson learned: These attacks highlight the high dimensionality of the data, which also makes them robust/scalable.)
- Privacy attacks on microtargeted ads [18] (Lesson learned: The attacker can play an active role, and can choose to be part of the dataset (e.g., via creating fake user profiles), and/or have an influence on the creation of the dataset (e.g., by setting up appropriate microtargeted ads).)
- Privacy attacks on collaborative filtering [6] (Lesson learned: The attacker can observe how the system changes over time (or other dimensions).)

1.2 Key Privacy Regulations and Laws

We will give a brief overview of key privacy regulations and laws such as GDPR (General Data Protection Regulation) [26], which took effect in May, 2018 as well as earlier laws such as HIPAA (Health Insurance Portability and Accountability Act).

1.3 Differential Privacy: Definition and Properties

We will motivate the need for a rigorous privacy guarantee, and present the notion of differential privacy [8–10]. Differential privacy is a formal guarantee for preserving the privacy of any individual when releasing aggregate statistical information about a set of people. In a nutshell, the differential privacy definition requires that the probability distribution of the released results be nearly the same irrespective of whether an individual's data is included as part of the dataset. As a result, upon seeing a published statistic, an attacker would gain very little additional knowledge about any specific individual. We will also give different interpretations of differential privacy, and its properties.

1.4 Privacy Techniques in Practice: Challenges and Lessons Learned

The privacy attacks over the last two decades have highlighted the need for adopting rigorous privacy techniques and demonstrated that it is highly non-trivial to balance the trade-offs between utility and privacy. Utility vs. privacy trade-offs have been studied in the literature, but handling the trade-offs between computation and communication resources is unique to industrial deployment. Current industrial deployments are highly distributed. Each device (e.g., a cell phone) holds a single data point, and the server computes aggregates over differentially private information obtained from the devices. Although server computation is cheap, client computation is expensive since the clients are usually low-power devices. Furthermore, client / server communication is expensive, and hence both the amount of communication and the number of rounds of communication have to be minimized.

Lessons learned: The notion of differential privacy is a principled foundation for privacy-preserving data analyses, which has been witnessing practical adoption in industry as well as in government (e.g., by U.S. Census Bureau [14, 19]). In particular, local differential privacy is a powerful technique suitable for internet-scale telemetry. While applying these techniques in practice, we have learned that resource constraints drive the algorithmic design as much as the

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY-NC-ND 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320085

privacy vs. utility trade-offs. In fact, often these constraints enable the design of theoretically optimal algorithms (e.g., [3]).

1.5 Case Studies in Industry

As part of the tutorial, we will also focus on the application of privacy-preserving data mining techniques in practice, by presenting case studies from different technology companies (including Google, Apple, LinkedIn, and Microsoft).

1.5.1 Case Study: Google's RAPPOR. We will describe Google's RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response) project [11, 12], which has been the first large scale deployment of differential privacy in industry. RAPPOR is built on the concept of randomized response, and enables a privacy-preserving way to learn software statistics (e.g., about how unwanted software is hijacking users' settings in Chrome browser) that can be used to improve browser security, find bugs, and provide better overall user experience. RAPPOR has been made available as an open-source project (https://github.com/google/rappor).

1.5.2 Case Study: Apple's Local Differential Privacy Deployment for iOS / macOS. We will present an overview of Apple's local differential privacy deployment for iOS and macOS, highlighting the application scenarios and the design choices [1].

Algorithm Design: We will describe the detailed algorithm based on the following patents (https://www.google.com/patents/US9594741 and https://www.google.com/patents/US9705908), and the follow-up paper [3]. The high-level idea is to use techniques from sketching and data streaming literature and adapt them to ensure differential privacy. A salient feature of these algorithms is that they provide optimal algorithms (in terms of error, storage, computation and communication) for locally differentially private heavy hitters.

Outreach: The deployed system currently runs on all iOS and macOS devices. The technology has been used for applications such as learning new words from user keyboards, learning health analytics, and device telemetry.

1.5.3 Case Study: Privacy-preserving Data Mining and Analytics at LinkedIn. We will next highlight the privacy challenges encountered during the design of LinkedIn Salary, a web-scale crowdsourcing system for secure collection and presentation of compensation insights to job seekers [15]. We will describe the privacy mechanisms based on techniques such as encryption, access control, de-identification, aggregation, and thresholding, and present open research challenges in the context of providing rigorous privacy guarantees [16].

We will also present the experiences from deploying differential privacy inspired mechanism for privacy-preserving analytics at LinkedIn [17]. The goal is to compute robust, reliable analytics in a privacy-preserving manner, while satisfying product requirements such as coverage, utility, and consistency. The key idea in this system is to use deterministic pseudorandom noise generation and perform post-processing to achieve data consistency.

1.5.4 Case Study: Microsoft's Local Differential Privacy Deployment for Collecting Windows Telemetry. We will briefly present the mechanisms underlying Microsoft's local differential privacy implementation, which have been deployed across millions of Windows devices to collect application usage statistics in a privacy-preserving manner [7].

1.6 Emerging Topics and Research Directions

In the last part of the tutorial, we will discuss emerging topics pertaining to privacy-preserving data mining and machine learning such as distributed private machine learning [5, 21–23], encode-shuffle-analyze architecture [4], and privacy amplification [13]. Finally, we will discuss open challenges and research directions for the community.

REFERENCES

[1] 2017. Learning with Privacy at Scale. *Apple's Machine Learning Journal* 1, 8 (2017).
[2] Michael Barbaro, Tom Zeller, and Saul Hansell. August 2006. A face is exposed for AOL searcher no. 4417749. *New York Times* (August 2006).
[3] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. 2017. Practical locally private heavy hitters. In *NIPS*.
[4] Andrea Bittau, Ulfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. 2017. PROCHLO: Strong privacy for analytics in the crowd. In *SOSP*.
[5] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *CCS*.
[6] Joseph A Calandrino, Ann Kilzer, Arvind Narayanan, Edward W Felten, and Vitaly Shmatikov. 2011. "You Might Also Like:" Privacy Risks of Collaborative Filtering. In *IEEE Symposium on Security and Privacy*.
[7] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. In *NIPS*.
[8] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*.
[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*.
[10] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3–4 (2014).
[11] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *CCS*.
[12] Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. 2016. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies* 3 (2016).
[13] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. 2018. Privacy Amplification by Iteration. In *FOCS*.
[14] Samuel Haney, Ashwin Machanavajjhala, John M Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. 2017. Utility cost of formal privacy for releasing national employer-employee statistics. In *SIGMOD*.
[15] Krishnaram Kenthapadi, Stuart Ambler, Liang Zhang, and Deepak Agarwal. 2017. Bringing Salary Transparency to the World: Computing Robust Compensation Insights via LinkedIn Salary. In *CIKM*. Available at https://arxiv.org/abs/1703.09845.
[16] Krishnaram Kenthapadi, Ahsan Chudhary, and Stuart Ambler. 2017. LinkedIn Salary: A System for Secure Collection and Presentation of Structured Compensation Insights to Job Seekers. In *IEEE PAC*. Available at https://arxiv.org/abs/1705.06976.
[17] Krishnaram Kenthapadi and Thanh T. L. Tran. 2018. PriPeARL: A Framework for Privacy-preserving Analytics at LinkedIn. In *CIKM*.
[18] Aleksandra Korolova. 2011. Privacy Violations Using Microtargeted Ads: A Case Study. *J. Privacy and Confidentiality* 3, 1 (2011).
[19] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory meets practice on the map. In *ICDE*.
[20] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*.
[21] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*.
[22] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable Private Learning with PATE. In *ICLR*.
[23] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. 2017. Is interaction necessary for distributed private learning?. In *IEEE Symposium on Security and Privacy*.
[24] Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks.

[25] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002).

[26] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer.

Human Mobility from theory to practice: Data, Models and Applications

Filippo Simini
Department of Engineering Mathematics, University of Bristol, UK
f.simini@bristol.ac.uk

Gianni Barlacchi
Fondazione Bruno Kessler and University of Trento, Italy
Povo, Italy
gianni.barlacchi@gmail.com

Roberto Pellungrini
Department of Computer Science, University of Pisa
roberto.pellungrini@di.unipi.it

Luca Pappalardo
ISTI-CNR, Pisa, Italy
luca.pappalardo@isti.cnr.it

ABSTRACT

The inclusion of tracking technologies in personal devices opened the doors to the analysis of large sets of mobility data like GPS traces and call detail records. This tutorial presents an overview of both modeling principles of human mobility and machine learning models applicable to specific problems. We review the state of the art of five main aspects in human mobility: (1) human mobility data landscape; (2) key measures of individual and collective mobility; (3) generative models at the level of individual, population and mixture of the two; (4) next location prediction algorithms; (5) applications for social good. For each aspect, we show experiments and simulations using the Python library "scikit-mobility" developed by the presenters of the tutorial.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → *Modeling methodologies.*

KEYWORDS

Human Mobility; Artificial Intelligence; Data Science; Generative Models; Predictive Algorithms

ACM Reference Format:
Filippo Simini, Gianni Barlacchi, Roberto Pellungrini, and Luca Pappalardo. 2019. Human Mobility from theory to practice: Data, Models and Applications. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320099

1 TOPIC AND RELEVANCE

The availability of geo-spatial mobility data (e.g., GPS traces, mobile phone records, social media records) is a trend that will grow in the near future. In particular, this will happen when the shift from traditional vehicles to autonomous, self-driving, vehicles, will change individual and public transportation, transforming our society, the economy and the environment. For this reason, understanding and modeling human mobility is of paramount importance for many present and future applications such as traffic forecasting, urban planning, estimating migratory flows, and epidemic modeling [12]. In this tutorial we will present a concise and intuitive overview on both the fundamental modeling principles of human mobility and artificial intelligence models applicable to specific mobility-related problems[1]. Starting from the general laws that govern human mobility, we will drive the audience through the main models for human mobility highlighting the parallelism between statistical and deep learning models, presenting the recent advances of the latter that are nowadays representing the state-of-the-art in many human mobility tasks, like next location prediction. To this end, we will review the state of the art of five aspects:

(1) **The human mobility data landscape**
A natural starting point is to describe the nature of empirical data which has been used in mobility research. In this part, we outline the main data sources available for mobility research and the relevant information that can be extracted from them. [3, 4, 6, 17, 35]

(2) **Under the microscope: Measuring individual and collective mobility patterns**
In this part, we will review some of the fundamental metrics and representations used to characterize human mobility, such as trip distance [8, 13], radius of gyration [13, 24, 25], mobility entropy [19, 34], origin-destination matrix [7], mobility motifs [30], and more.

(3) **Agents on the move: simulating mobility patterns**
This part will review the state of the art for generative models at both the individual level (i.e., generation of individual spatio-temporal trajectories) [2, 15, 23–25, 33] and the population level (i.e., generation of mobility flows) [16, 31, 32, 39].

(4) **Where's next? AI for human mobility**
After a short review of various artificial intelligence and machine learning models for human mobility [21, 22, 36, 38, 43] we will review recent advances based on deep learning, with particular focus on next location prediction [10, 18, 40, 41, 44].

(5) **Human mobility for Social Good and future challenges**
In this part, we will show how developing accurate predictive and generative mobility models can be greatly beneficial for several aspects of social good, from mobility in emergency

[1]The online version of the tutorial and all the updated material can be found at https://humanmobility-tutorial.github.io/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320099

scenarios [14] to the prevention of epidemic diffusion [9], nowcast well-being [26] and even the design of more sustainable smart cities [11]. We will discuss about present and future challenges on mobility-related problems such as ride-sharing [29], automatic discovery of urban regions [37, 42], prediction of health from human displacements [1, 5] and traffic forecasting [20, 28]. Finally, we discuss privacy issues related to the analysis of human mobility data [27].

ACKNOWLEDGMENTS

This work is funded by EU project SoBigData RI, grant #654024.

REFERENCES

[1] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS* 106, 51 (2009), 21484–21489.

[2] Hugo Barbosa, Fernando B. de Lima-Neto, Alexandre Evsukoff, and Ronaldo Menezes. 2015. The effect of recency to human mobility. *EPJ Data Science* 4, 1 (2015), 21. https://doi.org/10.1140/epjds/s13688-015-0059-8

[3] Hugo Barbosa and et al. 2018. Human mobility: Models and applications. *Physics Reports* (2018).

[4] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data* 2 (2015), 150055.

[5] Gianni Barlacchi, Christos Perentis, Abhinav Mehrotra, Mirco Musolesi, and Bruno Lepri. 2017. Are you getting sick? Predicting influenza-like symptoms using human mobility behaviors. *EPJ Data Science* 6, 1 (2017), 27.

[6] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4, 1 (2015), 10.

[7] Patrick Bonnel, Etienne Hombourger, Ana-Maria Olteanu-Raimond, and Zbigniew Smoreda. 2015. Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations. *Transportation Research Procedia* 11 (2015), 381 – 398. https://doi.org/10.1016/j.trpro.2015.12.032

[8] D. Brockmann, L. Hufnagel, and T. Geisel. 2006. The scaling laws of human travel. *Nature* 439, 7075 (26 01 2006), 462–465. http://dx.doi.org/10.1038/nature04292

[9] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences* 103, 7 (2006), 2015–2020.

[10] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. DeepMove: Predicting Human Mobility with Attentional Recurrent Networks. (2018).

[11] Michele Ferretti, Gianni Barlacchi, Luca Pappalardo, Lorenzo Lucchini, and Bruno Lepri. 2018. Weak nodes detection in urban transport systems: Planning for resilience in Singapore. *arXiv preprint arXiv:1809.07839* (2018).

[12] F. Giannotti, L. Pappalardo, D. Pedreschi, and D. Wang. 2013. *A Complexity Science Perspective on Human Mobility.* Cambridge University Press, 297âĂŞ314. https://doi.org/10.1017/CBO9781139128926.016

[13] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779.

[14] Renhe Jiang, Xuan Song, Zipei Fan, Tianqi Xia, Quanjun Chen, Satoshi Miyazawa, and Ryosuke Shibasaki. 2018. DeepUrbanMomentum: An Online Deep-Learning System for Short-Term Urban Mobility Prediction.. In *AAAI*.

[15] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. GonzÃąlez. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *PNAS* 113, 37 (2016). https://doi.org/10.1073/pnas.1524261113 arXiv:http://www.pnas.org/content/113/37/E5370.full.pdf

[16] Chaogui Kang, Yu Liu, Diansheng Guo, and Kun Qin. 2015. A Generalized Radiation Model for Human Mobility: Spatial Scale, Searching Direction and Trip Constraint. *PLoS ONE* 10, 11 (11 2015), 1–11. https://doi.org/10.1371/journal.pone.0143500

[17] Miao Lin and Wen-Jing Hsu. 2014. Mining GPS data for mobility patterns: A survey. *Pervasive and Mobile Computing* 12 (2014).

[18] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts.. In *AAAI*. 194–200.

[19] Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. 2013. Approaching the Limit of Predictability in Human Mobility. *Scientific Reports* 3 (11 10 2013).

[20] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. 2015. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems* 16, 2 (2015), 865–873.

[21] Wesley Mathew, Ruben Raposo, and Bruno Martins. 2012. Predicting future locations with hidden Markov models. In *Procs of the 2012 ACM conference on ubiquitous computing*. 911–918.

[22] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. 2009. Wherenext: a location predictor on trajectory pattern mining. In *Procs of the 15th ACM SIGKDD Int Conf on Knowledge discovery and data mining*. 637–646.

[23] Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. 2016. Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model. *Procedia Computer Science* 83 (2016), 934–939. https://doi.org/10.1016/j.procs.2016.04.188

[24] Luca Pappalardo and Filippo Simini. 2017. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery* (2017), 1–43.

[25] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-Laszlo Barabasi. 2015. Returners and explorers dichotomy in human mobility. *Nature Comm* 6 (2015).

[26] Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. 2016. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics* 2, 1 (01 Dec 2016), 75–92. https://doi.org/10.1007/s41060-016-0013-2

[27] Roberto Pellungrini, Luca Pappalardo, Francesca Pratesi, and Anna Monreale. 2017. A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. *ACM Trans. Intell. Syst. Technol.* 9, 3, Article 31 (Dec. 2017), 27 pages. https://doi.org/10.1145/3106774

[28] Nicholas G Polson and Vadim O Sokolov. 2017. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies* 79 (2017), 1–17.

[29] Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H Strogatz, and Carlo Ratti. 2014. Quantifying the benefits of vehicle pooling with shareability networks. *PNAS* 111, 37 (2014).

[30] Christian M. Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C. González. 2013. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* 10, 84 (2013). https://doi.org/10.1098/rsif.2013.0246 arXiv:http://rsif.royalsocietypublishing.org/content/10/84/20130246.full.pdf

[31] Filippo Simini, Marta C. Gonzalez, Amos Maritan, and Albert-Laszlo Barabasi. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (05 04 2012), 96–100. http://dx.doi.org/10.1038/nature10856

[32] Filippo Simini, Amos Maritan, and Zoltán Néda. 2013. Human mobility in a continuum approach. *PloS one* 8, 3 (2013), e60069.

[33] Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6, 10 (10 2010), 818–823.

[34] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021. https://doi.org/10.1126/science.1177170 arXiv:http://science.sciencemag.org/content/327/5968/1018.full.pdf

[35] Laura Spinsanti, Michele Berlingerio, and Luca Pappalardo. 2013. Mobility and Geo-Social Networks. In *Mobility Data: Modeling, Management, and Understanding*. 315–333.

[36] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. 2011. Human mobility, social ties, and link prediction. In *Procs of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*. 1100–1108. https://doi.org/10.1145/2020408.2020581

[37] Hongjian Wang and Zhenhui Li. 2017. Region Representation Learning via Mobility Flow. In *Procs of the 2017 ACM on Conference on Information and Knowledge Management*. 237–246.

[38] R. Wu, G. Luo, J. Shao, L. Tian, and C. Peng. 2018. Location prediction on trajectory data: A review. *Big Data Mining and Analytics* 1, 2 (June 2018), 108–127. https://doi.org/10.26599/BDMA.2018.9020010

[39] Xiao-Yong Yan, Chen Zhao, Ying Fan, Zengru Di, and Wen-Xu Wang. 2014. Universal predictability of mobility patterns in cities. *Journal of The Royal Society Interface* 11, 100 (2014), 20140834.

[40] Cheng Yang, Maosong Sun, Wayne Xin Zhao, Zhiyuan Liu, and Edward Y Chang. 2017. A Neural Network Approach to Jointly Modeling Social Networks and Mobile Trajectories. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 36.

[41] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Procs of the 26th Int Conf on World Wide Web*. 351–360.

[42] Nicholas Jing Yuan, Yu Zheng, and Xing Xie. 2018. Discovering Functional Zones in a City Using Human Movements and Points of Interest. In *Spatial Analysis and Location Modeling in Urban and Regional Systems*. Springer, 33–62.

[43] Chao Zhang, Keyang Zhang, Quan Yuan, Luming Zhang, Tim Hanratty, and Jiawei Han. 2016. Gmove: Group-level mobility modeling using geo-tagged social media. In *Procs of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*. 1305–1314.

[44] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction.. In *AAAI*. 1655–1661.

From Research Articles to Knowledge Graphs

Methods for Ontology-driven Knowledge Base Creation from Text

Vayianos Pertsas
Department of Informatics
Athens University of Economics and Business, vpertsas@aueb.gr

Panos Constantopoulos
Department of Informatics
Athens University of Economics and Business; Digital Curation Unit, Athena Research Centre,
panosc@aueb.gr

Abstract

Understanding and extracting knowledge contained in text and encoding it as linked data for the WEB is a highly complex task that poses several challenges, requiring expertise from different fields such as conceptual modeling, natural language processing and web technologies including web mining, linked data generation and publishing, etc. When it comes to the scholarly domain, the transformation of human readable research articles into machine comprehensible knowledge bases is considered of high importance and necessity today due to the explosion of scientific publications in every major discipline, that makes it increasingly difficult for experts to maintain an overview of their domain or relate ideas from different domains. This situation could be significantly alleviated by knowledge bases capable of supporting queries such as: find all papers that address a given problem; how was the problem solved; which methods are employed by whom in addressing particular tasks; etc. that currently cannot be addressed by commonly used search engines such as Google Scholar or Semantic Scholar.

This tutorial addresses the above challenge by introducing the participants to methods required in order to model knowledge regarding a given domain, extract information from available texts using advanced machine learning techniques, associate it with other information mined from the web in order to infer new knowledge and republish everything as linked open data on the Web. To this end, we will use a specific use case – that of the scholarly domain, and will show how to model research processes, extract them from research articles, associate them with contextual information from article metadata and other linked repositories and create knowledge bases available as linked data. Our aim is to show how methodologies from different computer science fields, namely natural language processing, machine learning and conceptual modeling, can be combined with Web technologies in a single meaningful workflow.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13-17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320090

CCS CONCEPTS

• Computing methodologies~Information extraction • Computing methodologies~Ontology engineering • Computing methodologies~Learning paradigms

Author Keywords

Conceptual Modeling; Ontology Engineering; Ontology Population; Information Extraction from Text; Machine Learning Methodologies; Linked Data

ACM Reference format:

Vayianos Pertsas and Panos Constantopoulos. 2019. From Research Articles to Knowledge Graphs: Methods for Ontology-driven Knowledge Base Creation from Text. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3320090

1 Topic and Relevance:

In this tutorial we will present a variety of methods covering the entire workflow of transforming human readable text (in particular that of research articles) into machine comprehensible linked data available on the Web. To this end we will: 1) introduce the participants to essential instruments for knowledge organization such as taxonomies and ontologies and will show how these can be used in a specific use case -that of modeling scholarly work- and 2) tutor the participants into using various methods for ontology population / knowledge base creation -based on our use case scenario- that cover mining information from publication metadata, extracting information from text using advanced machine learning techniques and ontology-based guidance, and combining those with mined resources from linked data repositories such as DBpedia and ORCID to infer new knowledge. Both parts will include a theoretical overview to explain the concepts and methods presented, as well as a presentation of the actual implementations in the form of an ontology for the first part and a system implemented in Python for the second part.

Transforming human readable text into knowledge graphs consisting of linked data available on the web is an active

research topic of high relevance to the WWW community since it requires bridging different Web technologies with methods from other computer science fields such as machine learning, natural language processing and conceptual modeling. The use case of modeling research processes and creating knowledge bases from research articles underlines the importance of the task, given the constant increase in the production and publishing of scientific papers in every discipline. Assisting researchers in answering queries that are currently beyond the capabilities of standard search engines, such as Google Scholar, Scopus or Semantic Scholar, which mostly navigate through author or citation graphs, can improve access to scientific literature and enhance researcher productivity. The construction of knowledge bases capable of answering complex queries of the form "*who* has done *what, when, why, how* and with *what* results" requires encoding information about research processes in a manner that enables reasoning and guides the extraction of information from publications, as well as the direct documentation of research activities. To this end, the level of expertise (with parts of our work published in journals and conferences such as IJDL [1], TPDL [2] , and ISWC [3]) as well as the long experience of the tutors on the subject, along with the scope of the tutorial, covering the state of the art in methodologies from knowledge representation and organization to information extraction from text and knowledge base creation, qualifies for a high-quality introduction to the topic.

2 Duration:

The proposed duration of the tutorial is half day, divided into two parts, where the exact division of the content will be adapted on demand to the participants. The following subjects will be covered:

First Part:

- Elements of knowledge organization (Controlled vocabularies, Taxonomies, Ontologies)
- Elements of Linked Data and Semantic Web Data Models (RDF, RDFS, OWL)
- Methodology and design patterns for modeling a specific domain. Use case: Scholarly Ontology for modeling scholarly work

Second Part:

- Knowledge base creation via ontology population. Use case: Populating the core concepts of Scholarly Ontology
- Distant supervision techniques for training Named Entity Recognizers of non-common type
- Information extraction from text using syntactic analysis and ontology-based semantics
- Entity and Relation Extraction from text using advanced Machine Learning techniques: feature engineering, embeddings, token/text representation.
- Metadata association and Linked Data integration and generation

3 Interaction Style:

Participants will be introduced to the methods and concepts covered in this tutorial through thorough presentations followed by Q&As and discussion on the practicalities of each covered subject. No specific equipment is required for attending the tutorial.

4 Audience:

Our target audience includes researchers and practitioners in web technologies, machine learning and information modeling with an interest in natural language processing and linked data methodologies. A background in computer science and some general familiarity with relevant concepts (ontologies, taxonomies, distant supervision, entity and relation extraction from text, linked data generation and integration as well as machine learning methods) are welcome but not necessary, as a brief overview of the employed concepts and methods will be included in the tutorial. Our aim is not to delve into the details of those concepts and methods, but to show how to use them and how they can be integrated into a meaningful workflow.

We expect the audience to take away a substantial overview of the state-of-the art in methodologies covering the entire workflow from designing models that capture knowledge in a specific domain (in our case that will be the domain of scholarly work), to the creation of knowledge bases by populating these models with information extracted from text (using machine learning techniques) along with linked data or publication metadata mined from the web.

5 Previous Editions:

This tutorial has never been presented in its entirety before (both parts covering conceptual modeling AND knowledge base creation). However, variations of the first part (conceptual modeling of research practices in the Digital Humanities) have been presented in the past in workshops with high attendance (more than 30 participants).

6 Tutorial Material:

Material will include tutorial notes, RDF serializations of the conceptual models, trained machine learning classifiers and samples of the actual produced knowledge bases (in the form of RDF triples), to be provided in time. No copyright issues are involved.

7 Equipment:

No additional equipment is needed for this tutorial.

8 Organizers:

Vayianos Pertsas holds a Dipl. Eng. Degree in Electrical and Computer Engineering from the University of Patras and a PhD degree in Informatics from the Athens University of Economics and Business. His research interests evolve around conceptual modeling and ontology population using information extraction

techniques that mainly focus on leveraging linked data and its applications, along with NLP and machine learning techniques. He has worked in the development of the NeDiMAH Methods Ontology in the ESF-funded Network for Digital Methods in the Arts and Humanities and in its evolution and operationalization in the form of the Scholarly Ontology. He has authored papers appearing in IJDL, ISWC, TPDL and co-tutored various workshops.

Panos Constantopoulos is Professor in the Department of Informatics, Director of the MSc Programme in Digital Methods for the Humanities, and former Dean of the School of Information Sciences and Technology, Athens University of Economics and Business. He is also affiliated with the Information Management Systems Institute of the "Athena" Research Centre, where he heads the Digital Curation Unit. He has previously been Professor and Chairman in the Department of Computer Science, University of Crete (1986-2003). From 1992 to 2003 he was head of the Information Systems Laboratory and the Centre for Cultural Informatics at the Institute of Computer Science, Foundation for Research and Technology - Hellas.

His scientific interests include: knowledge representation and conceptual modelling, ontology engineering, semantic information access, process mining, knowledge management systems, decision support systems, cultural informatics, digital libraries, digital curation and preservation.

He holds a Diploma in Electrical and Mechanical Engineering from the National Technical University of Athens (1978), a Master of Science in Electrical Engineering from Carnegie-Mellon University (1979) and a Doctor of Science in Operations Research from Massachusetts Institute of Technology (1983).

He has been principal investigator or scientific responsible on the part of his affiliation in 40 national or international research and development projects, in 13 of which project coordinator. He is currently heading "APOLLONIS-Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation", a three year, 4M Euro project jointly advancing the Greek components of CLARIN and DARIAH.

He has published over 100 articles in scientific journals, conference proceedings or as book chapters.

References

[1] Pertsas, V., Constantopoulos, P.: Scholarly Ontology: modelling scholarly practices. International Journal on Digital Libraries. 18, 173–190 (2017). doi:10.1007/s00799-016-0169-3

[2] Pertsas, V., Constantopoulos, P.: Ontology-driven information extraction from research publications. Lecture Notes in Computer Science. 11057 LNCS, 241–253 (2018). doi:10.1007/978-3-030-00066-0_21

[3] Pertsas, V., Constantopoulos, P.: Ontology-Driven Extraction of Research Processes. 1–16. doi:10.1007/978-3-030-00671-6_10

Tutorial: Sequence-Aware Recommender Systems

Massimo Quadrana
Pandora Media
Oakland, USA
mquadrana@pandora.com

Dietmar Jannach
Alpen-Adria-Universität Klagenfurt
Klagenfurt, Austria
dietmar.jannach@aau.at

Paolo Cremonesi
Politecnico di Milano
Milan, Italy
paolo.cremonesi@polimi.it

ABSTRACT

Recommender systems are widely used in online applications to help users find items of interest and help them deal with information overload. In this tutorial, we discuss the class of sequence-aware recommender systems. Differently from the traditional problem formulation based on a user-item rating matrix, the input to such systems is a sequence of logged user interactions. Likewise, sequence-aware recommender systems implement alternative computational tasks, such as predicting the next items a user will be interested in an ongoing session or creating entire sequences of items to present to the user. We propose a problem formulation, sketch a number of computational tasks, review existing algorithmic approaches, and finally discuss evaluation aspects of sequence-aware recommender systems.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Collaborative filtering**.

KEYWORDS

Recommender Systems; Sequence-Aware; Session-Based

ACM Reference Format:
Massimo Quadrana, Dietmar Jannach, and Paolo Cremonesi. 2019. Tutorial: Sequence-Aware Recommender Systems. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3320091

1 BACKGROUND

Recommender systems are omnipresent in the online world. They can be found on almost all major e-commerce, social media, and e-commerce sites and help user find relevant information in times of information overload. During the last twenty years, the dominant problem abstraction in academia is that of a user-item rating (or interaction) matrix, where the goal is to predict the relevance of each unseen items for a given user. While this abstraction has shown to be helpful in different ways, it also has a number of limitations [1]. For example, when timestamps of interactions are available, they are typically considered only within special "time-aware" algorithms. Furthermore, the user-item rating matrix abstraction does not support the representation of multiple interactions of one user with a given item over time.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320091

In real-world environments, it is probably not uncommon to keep track of *all* interactions of a user with a website, resulting in a sequence of user actions, including, for example, purchases, streaming activities, or "like" statements. Research on how to leverage this much richer information is however comparably sparse. In the tutorial, we discuss the family of sequence-aware recommender systems [2], which are based on such sequential logs and on a richer set of recorded user interactions, thereby supporting a number of computational tasks that are highly relevant in practice.

2 TUTORIAL CONTENT

After providing examples where the user-item rating matrix abstraction has its limitations in practice, we discuss an application-independent characterization of sequence-aware recommendation problems in terms of inputs, outputs and computational tasks, see also Figure 1.

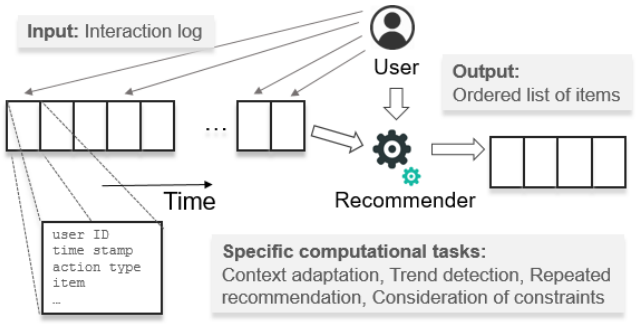


Figure 1: High-level overview of the Sequence-aware Recommendation problem, adapted from [2].

We then consider the "context adaptation" task in more detail and focus on the problem of *session-based* and *session-aware* recommendation, which received increased interest in recent years. Besides the discussion of selected algorithmic approaches designed, e.g., for the next-item prediction problem, we also look at today's research practice in terms of the academic evaluation of such algorithms. A practical demonstration complements the presentations of the tutorial.

REFERENCES

[1] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender Systems - Beyond Matrix Completion. *Commun. ACM* 59, 11 (2016), 94–102. DOI: https://doi.org/10.1145/2891406
[2] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *ACM Comput. Surv.* 51, 4 (2018). http://doi.acm.org/10.1145/3190616

Scalable Subgraph Counting: The Methods Behind The Madness

C. Seshadhri
University of California, Santa Cruz, CA
sesh@ucsc.edu

Srikanta Tirthapura
Iowa State University, Ames, IA
snt@iastate.edu

ABSTRACT

Subgraph counting is a fundamental problem in graph analysis that finds use in a wide array of applications. The basic problem is to count or approximate the occurrences of a small subgraph (the pattern) in a large graph (the dataset). Subgraph counting is a computationally challenging problem, and the last few years have seen a rich literature develop around scalable solutions for it. However, these results have thus far appeared as a disconnected set of ideas that are applied separately by different research groups. We observe that there are a few common algorithmic building blocks that most subgraph counting results build on. In this tutorial, we attempt to summarize current methods through distilling these basic algorithmic building blocks. The tutorial will also cover methods for subgraph analysis on "big data" computational models such as the streaming model and models of parallel and distributed computation.

CCS CONCEPTS

• **Information systems → Graph-based database models**; • **Theory of computation → Graph algorithms analysis**.

KEYWORDS

subgraph counting; motif counting; graphlet counting; sampling; edge orientation

ACM Reference Format:
C. Seshadhri and Srikanta Tirthapura. 2019. **Scalable Subgraph Counting: The Methods Behind The Madness. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320092**

1 INTRODUCTION AND GOAL

This tutorial focuses on *subgraph counting* (also called motif counting or graphlet analysis), which is an umbrella term that refers to problems where we wish to count or approximate the number of occurrences of a (small) subgraph H in a large graph G. An example problem is triangle counting, where we seek to approximate the number of instances of a *triangle*, a complete subgraph on three vertices, in a graph G. The term "subgraph counting" also encompasses versions where vertices have attributes, edges have timestamps, and cases where we wish to get (local) counts for each vertex. Depending on the application, there may be a need for algorithms in different computational models (streaming, distributed, sublinear, etc.).

Subgraph counting has a wide variety of uses, including:

- Social network analysis: Classifying social behavior, social behavior in gaming networks, explaining roles of nodes, predicting social tie strength, analyzing collaboration patterns. Subgraph counts form the basis of the widely used definition of graph clustering coefficients.
- Graph processing: Detecting dense subgraphs and communities, spam detection, modeling real-world graphs, semantic user search.
- Bioinformatics: Characterizing networks and roles of nodes within them, through the concepts of network motifs and graphlets.

Recent years have seen a surge of interest in subgraph counting in large graphs, in areas such as web search, social and biological network analysis. There is a dizzying array of papers on subgraph counting/approximation that have appeared in traditional data mining venues (The Web Conference, KDD, SDM, WSDM, etc.), database venues (VLDB, SIGMOD, PODS, etc.), and traditional theoretical computer science venues (FOCS, STOC, SODA, etc.) Such a large body of literature is overwhelming to track even for an active researcher in the area. Moreover, it is challenging for a practitioner to understand which result is best suited for their specific subgraph mining task.

Goal: The authors, while following the subgraph counting literature, have observed a set of algorithmic building blocks that are common to most of the results. Some examples of these building blocks are: edge sparsification, color coding, using graph orientations, and path sampling. Significant ingenuity is required to adapt these ideas to the problem and computational model at hand. Nonetheless, these common algorithmic building blocks form the foundation of much of this literature. Our aim is to highlight these techniques, and classify results according to the techniques they build on. The specific goals of the tutorial are to:

- Present algorithmic building blocks of scalable subgraph counting.
- Explain research results in the context of the above building blocks, and survey the research landscape.

Tutorial Structure: The tutorial will begin with an introduction to the problem of subgraph counting, its importance and applications. The tutorial will then introduce the various algorithmic building blocks. through existing results, and discuss how these building blocks can be implemented in different computational models. Given this foundation, we will classify many important results in the literature by the tools they build upon. We will not attempt an exhaustive survey, but we hope to give a bird's-eye view

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320092

of the research landscape. We will provide examples to illustrate algorithmic ideas, and provide proof sketches as appropriate.

2 ORGANIZERS

C. Seshadhri is an Associate Professor of Computer Science at the University of California, Santa Cruz. His primary interest is in mathematical foundations of big data, especially modeling and algorithms. His work spans many areas: sublinear algorithms, graph algorithms, graph modeling, scalable computation, and data mining. Seshadhri has worked extensively on the topic of graph algorithms and data mining. Related to the topic of the tutorial, he has contributed several significant results. These include the technique of "wedge sampling" for subgraph counting, fast exact algorithms for counting small subgraphs, and state-of-the-art theoretical and practical algorithms for approximating counts of k-cliques.

Srikanta Tirthapura is the Kingland Professor of Data Analytics at Iowa State University. His research is centered on the foundations of large-scale data analysis, especially streaming and parallel algorithms applied to large-scale data. He also has interests in applications of data analytics to areas such as security, and has worked extensively on algorithms in the streaming and incremental models. On the topic of the tutorial, he has contributed ideas such as "neighborhood sampling" for sampling and counting subgraphs from a data stream, which has been implemented in state-of-the-art systems for subgraph analysis, as well as works on parallel streaming algorithms for subgraph counting.

3 DURATION, AUDIENCE, OUTCOMES

This is a half-day tutorial, whose expected audience is:

- Algorithm designers and practitioners interested in large graph analysis
- Researchers in the domains of social network analysis and network measurement
- Researchers interested in the use of randomized methods for big data

The only prerequisites required is basic familiarity with graphs and probability, and some knowledge of undergraduate graph algorithms (such as adjacency list representations, BFS, DFS, etc).

Expected Outcomes: The audience will get an overview of the main algorithmic ideas used in subgraph counting results. Practitioners will get guidance on which ideas or tools to use for their specific problems (and models of computation), and will gain an understanding of where to look within the literature. Researchers in the areas of subgraph counting, graph algorithms, and randomized methods will gain from a survey of the cutting-edge results in this area.

Concept to Code: Deep Learning for Fashion Recommendation

Omprakash Sonie
Flipkart
omprakash.s@flipkart.com

Muthusamy Chelliah
Flipkart
muthusamy.c@flipkart.com

Shamik Sural
IIT Kharagpur
shamik@cse.iitkgp.ac.in

ABSTRACT

Deep Learning has shown significant results in various domains. In this tutorial, we provide conceptual understanding of embedding methods, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs). We present fashion use case and apply these techniques for modeling image, text as well as sequence data to figure out user profiles, give personalized recommendations tailored to changing user taste and interest. Given the image of a fashion item, recommending complementary matches is a challenge. Users' taste evolves over time and depends on persona. Humans relate objects based on their appearance and non-visual factors of lifestyle merchandise which further complicate recommendation task. Composing outfits in addition necessitates constituent items to be compatible - similar in some but different in other aspects.

CCS CONCEPTS

• **Applied computing → Online shopping**.

ACM Reference Format:
Omprakash Sonie, Muthusamy Chelliah, and Shamik Sural. 2019. Concept to Code: Deep Learning for Fashion Recommendation. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320100

1 INTRODUCTION AND EMBEDDING

Embedding is one of the key components in deep learning based systems. Hence, we go deeper in explaining Embeddings [15] [10] for learning vector representation of products, users and other types of data. Will deep dive into Skip-gram technique (with necessary foundation) which will be used for hands-on. We summarize the key concepts of the tutorial and the coverage in the following sections.

2 BASIC DEEP NEURAL NETWORK ARCHITECTURES

Convolutional Neural Network (CNN): A CNN is a multi-layer network having convolutional layers. Each layer consists of a number of filters. A filter is applied to the various local sub-regions of the image giving rise to a feature map. A CNN requires fewer parameters as the features are directly extracted from content such as product image or product review text. In a deep multi-layer CNN, the different layers learn different image features [14]. A common technique is to extract features from an image of a product using a pre-trained CNN.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320100

Recurrent Neural Network (RNN): RNNs model sequential data, capture memory of the past and can be trained by unfolding and doing back propagation through time. However, simple RNNs often fail to capture long distance temporal dependencies. Long Short Term Memory network (LSTM) and Gated Recurrent Unit (GRU) have been proposed to address this issue and have performed very well [6]. They enable recommender systems to model content sequences or temporal dynamics. RNNs have been used for session-based recommendations. [20] uses LSTMs to capture temporal dependencies of users and items.

3 STYLE DETECTION

In this section we describe what style is, its importance and how it is learnt. Style prediction involves analyzing the degree to which a style is exhibited either for items within same category or across different categories and finding which style is general and which one indicates a person's specific style. o address representation and learn inter-dependencies between features of fashion item we need to harness large fashion image corpus and rich meta [9]. There is a need to model simultaneously visual appearances of product as well as their evolution over time to uncover the complex and evolving visual factors that people consider when evaluating products [4]. An item consists of both Style and Category. Style information indicates preferences of users and has a significant effect in visual recommendation. Conventional methods model only categorical information. The model learns style features for understanding user preference [11].

4 RECOMMENDATION

In this section we cover various recommendation techniques. People make choices based on which items are alternative to each other (such as two T-shirts) and which items are complementary (T-Shirts and matching jeans). This uncovering human notions of visual relationships of items is modeled as network inference problem on graphs of related images [14]. An effective recommendation is not just recommending one item but a set of items that fit user's personal preferences. Instead of suggesting individual items, the model suggests a set of personalized items [7]. Aesthetic features (e.g., clean design, elegant style etc.) play important role in user's decision making process and varies from user to user and by time. The models extracts aesthetic features and includes in recommender system [21]. Users are more likely to buy outfits recommended based on high Attractive Quotient. The model suggests outfit based on Attractive Quotient and removes the outfits which are judged unattractive and suggests partial replacement [1].

It may be noted that, unlike other kinds of recommender systems, in the context of outfit recommendation, there are a very large number of possible outfits that are available on any online shopping site although for each such item, the number of actual customers could only be a small fraction of the total number of

existing or potentially new customers. Also, customers are often less guided by the choice of other customers than by their own personal preferences. Further, unlike movies and books, different users can have different budgetary constraints and would prefer that the recommender system meets such constraints specified either explicitly or implicitly.

5 OUTFIT COMPATIBILITY

In this section we cover various outfit compatibility approaches. In the fashion domain, discovering items that are functionally complementary or visually compatible is important. The model is designed to handle a set of complex relationships among items, as well as high-dimensional and semantically complicated features [5]. Users need to match clothes to make a suitable outfit. The model integrates deep neural networks and rich fashion domain knowledge for recommending outfit [17]. For suggesting an item that matches items in a stylish outfit (a collection of fashion items) the model learns compatibility by considering a fashion outfit fit to be a sequence and each items as a step. The visual embedding of items is included in the model [3]. Outfit recommender system need to discriminate substitutable products (that are interchangeable) and complementary products (which might be purchased together by users). The model leverages not just direct paths but complex dependencies - represented as sequence of relations and items in a product graph [19].

6 MISCELLANEOUS WORK

In this section we cover other approaches for fashion recommendation. In search, learning to rank of items is traditionally done over few hand-constructed features from item's text. The model combines traditional features with visual semantic features from images for ranking [12]. To help designer, brands and consumers to comprehend importance of different attributes in an outfit composition we need to include explainability in the deep learning based models. The model recommends a ranked list of outfit composition with interpretable matching scores [2]. Users are unhappy with wrong product size in popular categories (e.g., apparel, shoes) which results in high return rate for supplier. The model determines how a product of certain size fits a customer based on physical true sizes, sizes that are learnt from past product purchase and returns data [16]. Fashion domain is primarily visual. Traditional systems recommend similar looking 'top' for a 'bottom' query. To provide diversified items the model learns and recommends items based on generated complementary items [8].

7 CODE WALKTHROUGH AND CONCLUSION

We walk-through the code for a fashion recommender system on e-commerce dataset, summarize these models, parameters and understand what is going on behind the scene with various visualizations [13] [18]. We will use Jupyter notebook with already executed code for walkthrough. Fashion recommendation is an intricate problem that needs out of the box algorithm design. This is more due to the fact that there are various factors that need to be estimated from user interaction. For example, whether the user wants complementary or substitutable items, items similar to or completely different from previous purchases, intends to go with the current trend or deviate from the same, etc. There are also difficulties arising out of the need for making the recommendation personalized. While some of these necessitate quite a bit of interaction with the user, at the same time, too much of interaction might put off the potential customer. Profitability and competitiveness considerations also add to the overall complexity of the system design.

REFERENCES

[1] D. Banerjee, N. Ganguly, S. Sural, and K. S. Rao. 2018. One for the Road: Recommending Male Street Attire. In *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 571–582.

[2] Z. Feng, Z. Yu, Y. Yang, Y. Jing, J. Jiang, and M. Song. 2018. Interpretable Partitioned Embedding for Customized Multi-item Fashion Outfit Composition. In *International Conference on Multimedia Retrieval (ICMR)*. 143–151.

[3] X. Han, Z. Wu, Y. Jiang, and L. S. Davis. 2017. Learning Fashion Compatibility with Bidirectional LSTMs. In *ACM International Conference on Multimedia (MM)*. 1078–1086.

[4] R. He and J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *International Conference on World Wide Web (WWW)*. 507–517.

[5] R. He, C. Packer, and J. McAuley. 2016. Learning Compatibility Across Categories for Heterogeneous Item Recommendation. In *International Conference on Data Mining (ICDM)*. 937–942.

[6] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[7] Y. Hu, X. Yi, and L. S. Davis. 2015. Collaborative Fashion Recommendation: A Functional Tensor Factorization Approach. In *ACM International Conference on Multimedia (MM)*. 129–138.

[8] C. P. Huynh, A. Ciptadi, A. Tyagi, and A. Agrawal. 2018. CRAFT: Complementary Recommendations Using Adversarial Feature Transformer. *CoRR* abs/1804.10871 (2018).

[9] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan. 2014. Large Scale Visual Recommendations from Street Fashion Images. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. 1925–1934.

[10] Q. Le and T. Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on International Conference on Machine Learning (ICML) (ICML'14)*, Vol. 32. II–1188–II–1196.

[11] Q. Liu, S. Wu, and L. Wang. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 841–844.

[12] C. Lynch, K. Aryafar, and J. Attenberg. 2016. Images Don'T Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. 541–548.

[13] L. van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)* 9 (2008), 2579–2605.

[14] J. McAuley, C. Targett, Q. Shi, and A. van den H. 2015. Image-Based Recommendations on Styles and Substitutes. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 43–52.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NIPS)*. 3111–3119.

[16] V. Sembium, R. Rastogi, L. Tekumalla, and A. Saroop. 2018. Bayesian Models for Product Size Recommendations. In *International Conference on World Wide Web (WWW)*. 679–687.

[17] X. Song, F. Feng, X. Han, X. Yang, W. Liu, and L. Nie. 2018. Neural Compatibility Modeling with Attentive Knowledge Distillation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 5–14.

[18] O. Sonie, S. Sarkar, and S. Kumar. 2018. Concept to code: learning distributed representation of heterogeneous sources for recommendation. In *The ACM Conference Series on Recommender Systems (RecSys)*. 531–532.

[19] Z. Wang, Z. Jiang, Z. Ren, J. Tang, and D. Yin. [n. d.]. A Path-constrained Framework for Discriminating Substitutable and Complementary Products in E-commerce. In *ACM International Conference on Web Search and Data Mining (WSDM)*. 619–627.

[20] C. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing. 2017. Recurrent Recommender Networks. In *ACM International Conference on Web Search and Data Mining (WSDM)*. 495–503.

[21] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin. 2018. Aesthetic-based Clothing Recommendation. In *International Conference on World Wide Web (WWW)*. 649–658.

Representation Learning on Networks: Theories, Algorithms, and Applications

Jie Tang
Tsinghua University
Beijing, China
jietang@tsinghua.edu.cn

Yuxiao Dong
Microsoft Research
Redmond, WA
yuxdong@microsoft.com

ABSTRACT

Network representation learning offers a revolutionary paradigm for mining and learning with network data. In this tutorial, we will give a systematic introduction for representation learning on networks. We will start the tutorial with industry examples from Alibaba, AMiner, Microsoft Academic, WeChat, and XueTangX to explain how network analysis and graph mining on the Web are benefiting from representation learning. Then we will comprehensively introduce both the history and recent advances on network representation learning, such as network embedding and graph neural networks. Uniquely, this tutorial aims to provide the audience with the underlying theories in network representation learning, as well as our experience in translating this line of research into real-world applications on the Web. Finally, we will release public datasets and benchmarks for open and reproducible network representation learning research. The tutorial accompanying page is at https://aminer.org/nrl_www2019.

CCS CONCEPTS

• **Information systems → Social networks**; • **Computing methodologies → Unsupervised learning**; **Learning latent representations**; *Knowledge representation and reasoning*.

KEYWORDS

Network Embedding; Representation Learning; Graph Neural Networks; Feature Learning; Graph Mining; Network Science

ACM Reference Format:
Jie Tang and Yuxiao Dong. 2019. Representation Learning on Networks: Theories, Algorithms, and Applications. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320095

1 INTRODUCTION

In this tutorial, we aim to provide a comprehensive review for network representation learning. First, we will identify a wide range of important problems in networks, such as link prediction [13, 19], node attribute inference [5], and social influence modeling [21]. We then introduce the conventional network mining pipeline for addressing these problems [6]. Finally, we will show how the recent emergence of network representation learning offers a new paradigm for mining and learning with networks [2, 10].

With that, we will dive into the current development of network representation learning and introduce the important methods and milestones along each direction. The uniqueness of this tutorial lies in its emphasis on both the theoretical understanding and real applications of network representation learning. On one hand, we will try to provide the audience with the underlying theories in this topic and how different models can be unified and connected with each other. On the other hand, we will introduce how to translate the network representation learning research into online and mobile applications on Alibaba, AMiner, Microsoft Academic, MOOC, and Tencent Games and WeChat.

The main content of this tutorial is organized as follows. First, we introduce the history of (network) representation learning and how this discipline is developed from graph theory, deep learning, and natural language processing [1, 12, 14]. A brief timeline of network representation learning research is summarized in Figure 1, including both the development of network embedding and graph neural networks. Together with its history, we will also present the basic knowledge and motivation for learning network representations.

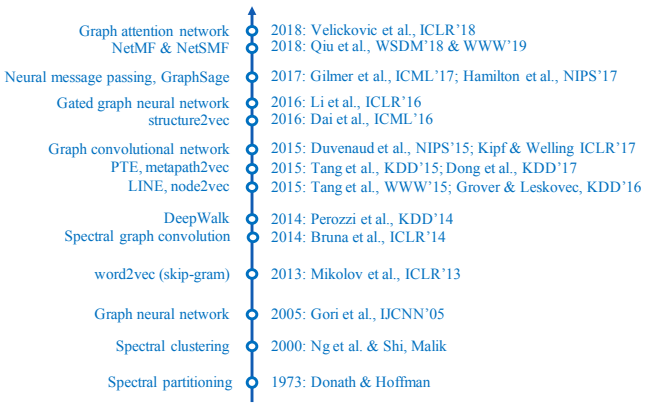


Figure 1: A timeline of network representation learning and graph neural network research as of 2018.

Second, we will introduce and categorize the recent developments of network representation learning. In specific, our focus will be on 1) skip-gram based algorithms (i.e., DeepWalk [15], LINE [20], node2vec [8]), and metapath2vec [4], 2) matrix factorization based models (i.e., NetMF [17] and NetSMF [16]), and 3) graph neural networks (i.e., Graph Convolution [3, 11], GraphSage [9], neural

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320095

message passing [7], and Graph Attention [23]). In addition, we will show 1) how to incorporate social and network theories into representation learning, and 2) how to leverage these techniques to handle more complex networks, such as dynamic and heterogeneous networks [4, 24].

Third, we will theoretically analyze the underlying mechanism of different types of network representation learning models and see how they can be connected with graph theory [16, 17]. For example, we will show that DeepWalk is actually implicitly factorizing a low-rank transformation of a network's normalized Laplacian matrix, and LINE, in theory, is a special case of DeepWalk.

In addition, we will also introduce how these network embedding techniques help real-world applications in Alibaba, AMiner, Microsoft Academic, Tencent Games, and WeChat. The covered applications include user profiling, behavior modeling, social influence, information propagation, and online recommendation.

Finally, we will summarize the problems and models introduced for network representation learning, and discuss the challenges and future directions. To facilitate open research for learning network representations, we will release large-scale public datasets and benchmarks during the tutorial. The to-be-released dataset for network representation learning is by far the largest Open Academic Graph composing of 200 million nodes and three billion links [18, 22].

The outline of the tutorial organization and arrangement is listed below. Each part is structured to introduce important problems, recent & advanced representation learning models, and direct applications from our collaborations with online/mobile social networking sites, which also forms as a unique value of this tutorial.

- I: Preliminaries for representation learning on networks
  - Real network problems in Alibaba, Tencent, etc.
  - Basic graph theories
  - Basic deep learning
  - Basic natural language processing (skip-gram)
  - Other related concepts
- II: Network representation learning algorithms
  - Skip-gram based network embedding models
  - Matrix factorization based embedding models
  - Graph neural networks
- III: Network embedding theory
  - Theories underlying skip-gram based models
  - Theories unifying factorization and graph convolution
  - Unified network embedding models
  - Fast embedding models
- IV: Network representation learning applications
  - Applications on Web user profiling
  - Applications on social influence on the Web
  - Applications on Web user behavior modeling
  - Applications on online recommendation
- V: Summary and open network embedding challenge
  - Review on algorithms and theories
  - Release the Open Benchmarks on Network Embedding
  - Release the Open Network Embedding Datasets.

In summary, this tutorial will provide an overview of network representation learning with the latest techniques and trends covered. One of its major goals is to generate theoretical insights into the rapid development of this field. The other one is to bridge network representation learning research with industry applications. Finally, to empower open and reproducible research, we will also release large-scale network datasets and benchmarks for representation learning.

REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 35, 8 (2013), 1798–1828.

[2] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017), 18–42.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral networks and locally connected networks on graphs. In *ICLR '14*.

[4] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In *KDD '17*.

[5] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. 2014. Inferring User Demographics and Social Strategies in Mobile Social Networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. ACM, 15–24.

[6] David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

[7] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1263–1272.

[8] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD '16*. ACM, 855–864.

[9] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NIPS'17*. 1024–1034.

[10] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *IEEE Data(base) Engineering Bulletin* 40 (2017), 52–74.

[11] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR '17*.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[13] David Liben-Nowell and Jon Kleinberg. 2003. The Link Prediction Problem for Social Networks. In *ACM CIKM '03*. 556–559.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop '13*.

[15] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *KDD '14*. ACM, 701–710.

[16] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Chi Wang, Kuansan Wang, and Jie Tang. 2019. NetSMF: Large-Scale Network Embedding as Sparse Matrix Factorization. In *WWW '19*. n/a.

[17] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In *WSDM '18*. ACM, 459–467.

[18] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW '15 Companion*. ACM, 243–246.

[19] Jie Tang, Tiancheng Lou, and Jon Kleinberg. 2012. Inferring social ties across heterogenous networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 743–752.

[20] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW '15*. 1067–1077.

[21] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. 2009. Social Influence Analysis in Large-scale Networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*. 807–816.

[22] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *KDD '08*. 990–998.

[23] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[24] Xian Wu, Baoxu Shi, Yuxiao Dong, Chao Huang, and Nitesh V Chawla. 2019. Neural Tensor Factorization for Temporal Interaction Learning. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 537–545.

Continuous Analytics of Web Streams

Half-Day Tutorial at The Web Conference 2019

Riccardo Tommasini
Politecnico di Milano
Milan, Italy
riccardo.tommasini@polimi.it

Robin Keskisärkkä
Linköping University
Linköping, Sweden
robin.keskisarkka@liu.se

Jean-Paul Calbimonte
Haute Ecole Spécialisée de Suisse Occidentale (HES-SO)
Sierre, Switzerland
jean-paul.calbimonte@hevs.ch

Eva Blomqvist
Linköping University
Linköping, Sweden
eva.blomqvist@liu.se

Emanuele Della Valle
Politecnico di Milano
Milan, Italy
emanuele.dellavalle@polimi.it

Albert Bifet
LTCI, Telecom ParisTech
Paris, France
albert.bifet@telecom-paristech.fr

ABSTRACT

This half-day tutorial provides a comprehensive introduction to web stream processing, including the fundamental stream reasoning concepts, as well as an introduction to practical implementations and how to use them in concrete web applications. To this extent, we intend to (1) survey existing research outcomes from Stream Reasoning / RDF Stream Processing that arise in querying, reasoning on and learning from a variety of highly dynamic data, (2) introduce deductive and inductive stream reasoning techniques as powerful tools to use when addressing a data-centric problem characterized both by variety and velocity, (3) present a relevant use-case, which requires to address data velocity and variety simultaneously on the web, and guide the participants in developing a web stream processing application.

CCS CONCEPTS

• **Information systems** → *Graph-based database models*; *Data streams*; *Stream management*; • **Theory of computation** → *Automated reasoning*.

KEYWORDS

Web Stream Processing; Stream Reasoning; RDF Stream Processing

ACM Reference Format:
Riccardo Tommasini, Robin Keskisärkkä, Jean-Paul Calbimonte, Eva Blomqvist, Emanuele Della Valle, and Albert Bifet. 2019. Continuous Analytics of Web Streams: Half-Day Tutorial at The Web Conference 2019. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3320088

1 TUTORIAL DESCRIPTION

Data streams are increasingly becoming available on the Web. A variety of sources give origin to data streams, including social networks, mobile phones, smart homes, healthcare devices, and parts of the modern infrastructure. From new opportunities arise new challenges. A common problem in scenarios involving data streams on the web is how data can be integrated to enable the creation of new knowledge, either top-down (using symbolic deductive reasoning) or bottom-up (using machine learning or inductive reasoning). Reasoning techniques are a possible solution to help with this integration. However, while modern deductive reasoners scale up in the static domain of ontological knowledge and machine learning techniques can learn from stationary data streams, inductive and deductive reasoning upon rapidly changing information has received attention only in the last decade [7]. Dealing with streams with concept drift, is nowadays considered one of the main challenges in machine learning, and more specifically, how to address it using accurate methods, as ensemble classifiers [8, 9]. The combination of reasoning techniques with data streams has given rise to the new research area of Stream Reasoning, i.e., reasoning on highly dynamic flows of information [4]. Stream Reasoning research area that has started to produce influential results relevant to both the Semantic Web [11] and stream processing communities [5].

Topic and Relevance: The goals of the tutorial are to (i) provide fundamental notions of processing streams on the web, (ii) offer an overview of current challenges and state of the art, (iii) describe the process of publication and description of data streams on the web, (iv) showcase different techniques and technologies for processing web streams (both in terms of machine learning and in terms of querying and deductive reasoning), and (v) outline the process of developing and deploying web-stream processing applications.

The contents of this tutorial is highly relevant for WWW 2019 attendees, as it focuses on practical aspects of developing and deploying applications that publish, produce, and process streaming (or rapidly changing) structured data. Web Stream processing technologies are now mature and reliable enough to build a tutorial with a stronger focus on hands-on sessions.

Interaction Style. This is a half-day tutorial that alternates lectures, demos, and hands-on sections. In particular, the tutorial offers to the audience:

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320088

- Lectures providing an overview of the use-cases and scenarios where web stream processing provides valuable advantages;
- demos about a subset of these technologies to reason over dynamic data, with techniques and tools developed by several research groups;
- practical experience through hands-on sessions were the audience can interact with existing tools (including but not limited to presenters' ones).

Audience. The tutorial targets researchers, knowledge workers, and practitioners interested in approaching the topic of web stream processing (both querying and reasoning) and who want to understand the current state-of-the-art as well as the future directions. The technologies and topics on this tutorial are relevant for people from IoT and sensor communities, as well as social media, pervasive health, oil industry, etc., who have to analyze in real-time massive amounts of streaming data.

2 DURATION & PROGRAM

This half-day tutorial aims at introducing different existing approaches for querying, analyzing, and reasoning over web streams. It also provides guidelines to developing and deploying Stream Reasoning applications. In the following, we provide a detailed program with the duration. Notably, we kept 1 hour as buffer.

(1) **Background for Web Stream Processing** [0.75h]
  (a) RDF & SPARQL in a nutshell [slides]
  (b) Stream Processing 101 [slides]
    (i) Paradigm shift to continuous semantics [slides]
    (ii) Stream processing languages and architectures [slides]
    (iii) Streaming Machine Learning [2]
(2) **Web Stream Processing: From Challenges to Opportunities** [0.75h] [slides]
  (a) The Stream Reasoning research question
  (b) Taming data variety and velocity on the web
  (c) Learning from unstationary streams
  (d) Use-cases & applications
(3) **Anatomy of Web Stream Processing Applications** [2h]
  (a) Publishing and describing streams on the web [0.75h]
    (i) Streams on the Web: publication challenges
    (ii) A vocabularies[1] and tools[2] for publishing web streams
    (iii) Demo 1: TripleWave for DBPedia changes
    (iv) Exercise 1: Using VoCaLS to create a stream catalog
  (b) **Processing Web Streams** [0.75h] [slides]
    (i) Stream Processing Languages in the Big Data Era [10]
    (ii) Stream Processing Languages in the Web Era [1, 3, 6, 12]
    (iii) Demo 2: representative RSP dialects and engines: C-SPARQL, CQELS, Yasper
    (iv) Exercise 2: Modeling web stream processing services with VoCaLS
    (v) Demo 3: Machine Learning using MOA[3]
(4) **Building a Web Streams Application** [2.5h] [materials]
  (a) Introduction of the tutorial running example
  (b) Designing the application: step-by-step
  (c) RSP Lab: An environment for analyzing streams and monitor performance[4]
  (d) Demo 3: Processing RDF Streams in the Wild
  (e) Exercises 3–4: Streaming Data Analytics
    (i) classification
    (ii) anomaly detection
  (f) Exercise 5–6: Streaming Data Management
    (i) stream discovery and exploration
    (ii) reasoning over streams
  (g) Benchmarking and Monitoring of Web Stream processing engines in RSP Lab [13].
(5) **Wrap-up and discussion** [0.25h]
  (a) On-going research trends, real-world deployments
  (b) Open problems and future directions

3 PREVIOUS EDITIONS

The tutorial follows from the Stream Reasoning for Linked Data (SR4LD) tutorial series, successfully held at ESWC 2011, SemTech 2011, ISWC 2013, ISWC 2014, ISWC 2015; the RDF Stream Processing (RSP) tutorial series at ESWC 2014 and ISWC 2016; the Stream Reasoning: Managing Velocity and Variety in Big Data tutorial at DEBS 2016; the tutorial on How to Build a Stream Reasoning Application co-located with ISWC 2017; and the Tutorial on Engineering Web Stream Processing applications at ICWE 2018, RW 2018; and the Stream Reasoning tutorial at KR 2018.

This tutorial differs from these previous events in that it emphasizes on the use of well known data-science tools such as research notebooks, and an enhanced experience for the data scientist, while using underlying stream reasoners, RSP engines, and tools for streaming machine learning. Links to previous tutorials and their complete slide sets are available in the hyperlinks above, and at http://streamreasoning.org/events.

4 TUTORIAL MATERIAL & EQUIPMENT

We provide links to teaching material, including slides and hands-on resources in Section 2, alongside with the tentative program.

Material and slides will be adapted for this tutorial by our organizing committee. Moreover, all the exercises make use of Jupyter Notebooks, which allows us to provide the infrastructure necessary for the tutorial exercises, while hiding some of the underlying complexities for the users. Jupyter Notebooks is one of the most popular tools for data science in both academia and industry and the winner of 2017 ACM Software System Award[5] Docker[6] images and a virtual machine will be provided for download before the tutorial day, to avoid configuration issues and to simplify deployment and execution of the exercises.

5 BIOGRAPHIES

5.1 Presenters

Emanuele Della Valle holds a PhD in Computer Science from the Vrije Universiteit Amsterdam and a Master degree in Computer Science and Engineering from Politecnico di Milano He is assistant

[1] https://w3id.org/rsp/vocals
[2] https://github.com/streamreasoning/TripleWave
[3] https://moa.cms.waikato.ac.nz/
[4] https://github.com/riccardotommasini/rsp-kernel
[5] https://awards.acm.org/software-system
[6] https://www.docker.com/

professor at the Department of Electronics, Information and Bioengineering of the Politecnico di Milano. In more than 15 years of research, his research interests covered Big Data, Stream Processing, Semantic technologies, Data Science, Web Information Retrieval, and Service Oriented Architectures. He started the Stream Reasoning research field [4] by focusing the research community's interest on this theme through a number of personal initiatives (workshops, projects, journal papers) that contributed to establishing Stream Reasoning as a recognized research and industrial sector. Starting from 2011, he organized more than 15 tutorials and workshops on Stream Reasoning[7]. His work on Stream Reasoning was applied in analyzing Social Media, Mobile Telecom and IoT data streams in collaboration with Telecom Italia, IBM, Siemens, Oracle, Indra, and Statoil. In 2015, he started up a company (Fluxedo) to commercialize the open source results of Stream Reasoning research.

Jean-Paul Calbimonte is a senior researcher at HES-SO Valais-Wallis, Switzerland. His work focuses on data integration and Semantic Web, applied to streaming sensor data sources. He has worked on ontology-based access for streaming data, resulting in the SPARQLStream language and the Morph-streams evaluator. He also helps coordinating the W3C Community Group on RDF Stream Processing (RSP). He has previously presented in the ISWC 2013 and ISWC 2014 Stream Reasoning for Linked Data Tutorial, the ISWC 2016 RDF Stream Processing tutorial, and the ESWC 2014 RDF Stream Processing (RSP) tutorial. He has also presented a tutorial on Linked Stream Data at the Winter School on Knowledge Technologies for Complex Business Environments (2011) and a tutorial on Semantic Sensor Web at the Extended Semantic Web Conference (ESWC 2011).

Riccardo Tommasini is a PhD student at the Department of Electronics and Information of the Politecnico di Milano. He enrolled in November 2015, focusing on how to realize an Efficient yet Expressive stream reasoning approach. His research interest comprise Reasoning and Ontology Based Data Access, Stream Processing and Complex Event Processing, Temporal Logics and Benchmarking. Riccardo's teaching activities comprise (i) a Big Data Tutorial with practical classes on Esper and C-SPARQL during his visiting at Kno.e.sis centre at Wright State University, Dayton, Ohio. (ii) He is a teaching assistant at Politecnico di Milano for the courses: Interoperability and Semantic Web Technologies (15-16) and Principles of Programming Languages (16-217, 17-18).

5.2 Organizers

Robin Keskisärkkä is PhD student at the Department of Computer and Information Science at Linköping University. His research focus is primarily on using Semantic Web technologies for Complex Event Processing, and on approaches for managing and propagating uncertainty in the context of streaming data. He is a member of the the W3C Community Group on RDF Stream Processing (RSP).

Eva Blomqvist is an Assistant Professor in the Department of Computer and Information Science at Linköping University. Her research focus is on ontologies an ontology engineering, and she was one of the researchers who initially proposed the notion of Ontology Design Patterns (ODPs). During her PhD she worked on semi-automatic ontology development, so-called ontology learning. Eva has also been actively involved in the development, refinement and evaluation of the eXtreme Design ontology engineering methodology, which was the first agile ontology engineering methodology when it was proposed in 2009. More recently Eva has been involved in several projects applying ontologies in various contexts, e.g. in decision support systems, and using ontologies and other Semantic Web technologies to perform semantic Complex Event Processing to make sense of streaming data.

Albert Bifet is Professor at Telecom ParisTech, Head of the Data, Intelligence, and Graphs (DIG) Group, and Honorary Research Associate at the WEKA Machine Learning Group at the University of Waikato. Previously he worked at Huawei Noah's Ark Lab in Hong Kong, Yahoo Labs in Barcelona, University of Waikato and UPC BarcelonaTech. He is the co-author of a book on Machine Learning from Data Streams. He is one of the leaders of MOA and Apache SAMOA software environments for implementing algorithms and running experiments for online learning from evolving data streams. He was serving as co-chair of the Industrial track of IEEE MDM 2016, ECML PKDD 2015, and as co-chair of BigMine (2017-2012), and ACM SAC Data Streams Track (2018-2012).

[7] http://streamreasoning.org/events

REFERENCES

[1] Davide Francesco Barbieri, Daniele Braga, Stefano Ceri, Emanuele Della Valle, and Michael Grossniklaus. 2010. C-SPARQL: a Continuous Query Language for RDF Data Streams. *Int. J. Semantic Computing* 4, 1 (2010), 3–25.

[2] Albert Bifet, Ricard Gavaldà, Geoff Holmes, and Bernhard Pfahringer. 2018. *Machine Learning for Data Streams with Practical Examples in MOA.* MIT Press. https://moa.cms.waikato.ac.nz/book/.

[3] Jean-Paul Calbimonte, Óscar Corcho, and Alasdair J. G. Gray. 2010. Enabling Ontology-Based Access to Streaming Data Sources. In *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I.* 96–111.

[4] Emanuele Della Valle, Stefano Ceri, Frank van Harmelen, and Dieter Fensel. 2009. It's a Streaming World! Reasoning upon Rapidly Changing Information. *IEEE Intelligent Systems* 24, 6 (2009), 83–89.

[5] Emanuele Della Valle, Daniele Dell'Aglio, and Alessandro Margara. 2016. Taming velocity and variety simultaneously in big data with stream reasoning: tutorial. In *DEBS.* ACM, 394–401.

[6] Daniele Dell'Aglio, Emanuele Della Valle, Jean-Paul Calbimonte, and Óscar Corcho. 2014. RSP-QL Semantics: A Unifying Query Model to Explain Heterogeneity of RDF Stream Processing Systems. *Int. J. Semantic Web Inf. Syst.* 10, 4 (2014), 17–44.

[7] Daniele Dell'Aglio, Emanuele Della Valle, Frank van Harmelen, and Abraham Bernstein. 2017. Stream reasoning: A survey and outlook. *Data Science* 1, 1-2 (2017), 59–83. https://doi.org/10.3233/DS-170006

[8] João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4 (2014), 44:1–44:37.

[9] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. 2017. A Survey on Ensemble Learning for Data Stream Classification. *ACM Comput. Surv.* 50, 2 (2017), 23:1–23:36.

[10] Martin Hirzel, Guillaume Baudart, Angela Bonifati, Emanuele Della Valle, Sherif Sakr, and Akrivi Vlachou. 2018. Stream Processing Languages in the Big Data Era. *SIGMOD Record* 47, 2 (2018), 29.

[11] Alessandro Margara, Jacopo Urbani, Frank van Harmelen, and Henri E. Bal. 2014. Streaming the Web: Reasoning over dynamic data. *J. Web Sem.* 25 (2014), 24–44. https://doi.org/10.1016/j.websem.2014.02.001

[12] Danh Le Phuoc, Minh Dao-Tran, Josiane Xavier Parreira, and Manfred Hauswirth. 2011. A Native and Adaptive Approach for Unified Processing of Linked Streams and Linked Data. In *The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I.*

[13] Riccardo Tommasini, Emanuele Della Valle, Marco Balduini, and Daniele Dell'Aglio. 2016. Heaven: a framework for systematic comparative research approach for RSP engines. In *13th Extended Semantic Web Conference, ESWC 2016, Heraklion, Crete, Greece.* 87–92.

Socially Responsible Natural Language Processing

Yulia Tsvetkov
Carnegie Mellon University
ytsvetko@cs.cmu.edu

Vinodkumar Prabhakaran
Google
vinodkpg@google.com

Rob Voigt
Stanford University
robvoigt@stanford.edu

ABSTRACT

As language technologies have become increasingly prevalent in analyzing online data, there is a growing awareness that decisions we make about our data, methods, and tools often have immense impact on people and societies. This tutorial will provide an overview of real-world applications of Natural Language Processing technologies and their potential ethical implications. We intend to provide the researchers with an overview of tools to ensure that the data, algorithms, and models that they build are socially responsible. These tools will include a checklist of common pitfalls that one should avoid, as well as methods to mitigate these issues. Issues of bias, ethics, and impact are often not clear-cut; this tutorial will also discuss the complexities inherent in this area.

CCS CONCEPTS

• **Social and professional topics** → **Codes of ethics**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

socially responsible computing; ethical considerations in computing; ethics in natural language processing; bias in data and models; textual data; privacy; human-generated data

ACM Reference Format:
Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2019. Socially Responsible Natural Language Processing. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308558.3320097

1 TOPIC AND RELEVANCE

In this tutorial we will discuss the philosophical foundations of ethical research along with state-of-the art machine learning fairness methods, with a particular focus on language data, algorithms, and applications. Major discussion topics include:

- Foundations of ethics in NLP: what is ethics, history, medical and psychological experiments, ethical decision making, a case study that identifies ethical pitfalls in working with computational models that can affect human lives.
- Working with people and human biases in language data: IRB and human subjects, working with crowdsourced workers, biases inherent in language data, privacy considerations.
- Methods to detect and mitigate biases in NLP models.

With a half-day tutorial and many possible topics, we do not plan an exhaustive treatment of this material. One central goal is to raise awareness for participants of the relevant issues, so that when they return to their research they will be more able to notice ways that ethical issues may play out in different contexts. To achieve this goal, we will aim for a "T-shape" in terms of breadth and depth: to briefly mention a number of core ethical questions, and then to drill down into a few particular case studies to see how these issues play out in real research settings.

As computational linguists, we often see the internet as a vast, interconnected set of "documents," which are largely composed of text written in natural language. Many WWW participants do work that touches upon working with language on the internet directly or indirectly, but even for those who do not work with language, issues of social bias are likely to find substantial relevance. One issue we hope to discuss with the WWW community in detail is the variety of sampling biases in internet research: often data accessibility, the availability of APIs, and the terms of service for web platforms influence what research is possible to do, and we would encourage a discussion of strategies to acknowledge and mitigate these issues.

2 AUDIENCE

We expect participants from a wide array of backgrounds, including researchers, engineers, and end users of WWW technologies. No prior experience with NLP/ML is required, but we believe that our tutorial will most benefit those who are currently using NLP or are intending to use NLP in the near future in their research/products.

3 PREVIOUS EDITIONS

The first iteration of this tutorial (in the half-day format) was given at NAACL 2018: https://sites.google.com/corp/view/srnlp/tutorial-naacl18. Link to the slides are at the bottom of the page. We have updated the content to include more recent work in this area.

4 RELEVANT LINKS AND COURSES

We maintain a web page relevant to the tutorial at https://sites.google.com/view/srnlp.

Additional relevant courses in the intersection of Ethics and NLP:

- Emily Bender at Univ. of Washington: http://faculty.washington.edu/ebender/2017_575/
- Graham Hirst at Univ. of Toronto: http://www.cs.utoronto.ca/~gh/cscD03/index.shtml
- Yulia Tsvetkov and Alan W Black's at CMU: http://demo.clab.cs.cmu.edu/ethical_nlp/

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3320097

The Challenge of API Management: API Strategies for Decentralized API Landscapes

Erik Wilde
Good API
erik.wilde@dret.net

Mike Amundsen
Amundsen Training
mca@amundsen.com

ABSTRACT

The Web as the world's largest information system has largely settled on a solid foundation of HTTP-based connectivity, and the representation of *User Interface (UI)* information resources through a mix of HTML and scripting. In contrast, the similarly rapidly evolving "Web of Services" is still based on a more diverse and more quickly evolving set of approaches and technologies. This can make architectural decisions harder when it comes to choosing on how to expose information and services through an *Application Programming Interface (API)*. This challenge becomes even more pronounced when organizations are faced with developing strategies for managing constantly growing and evolving API landscapes.

This tutorial takes participants through two different journeys. *The first one* is a journey discussing API styles and API technologies, comparing and contrasting them as a way to highlight the fact that there is no such thing as the one best choice. The goal of this first journey is to provide an overview of how APIs are used nowadays in research and in industry. *The second journey* discusses the question of how to define an API strategy, which focuses both on helping teams to make effective choices about APIs in a given context, and on how to manage that context over time when large organizations nowadays have thousands of APIs, which will continue to evolve constantly.

KEYWORDS

API; SOA; Web Services; Open Data

ACM Reference Format:
Erik Wilde and Mike Amundsen. 2019. The Challenge of API Management: API Strategies for Decentralized API Landscapes. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320089

1 TUTORIAL OVERVIEW

APIs are becoming an increasingly important part of the Web. In the same way as the network effect helps the human-facing Web to grow, this is true for the machine-facing Web. There is an ever-increasing set of services available, originating in global and explosively growing scenarios such as connected mobility, smart cities, or the Web of things. On the other hand, developers can build new and innovative applications all the time by connecting with these services, and aggregating them into new services, or into human-facing apps on the Web and other platforms. Even the trend towards native apps on mobile devices wouldn't be possible without the connectivity provided by the Web. The Web simply is much bigger than just the part directly interfacing with humans.

Digital Transformation and similar trends are encouraging organizations to increasingly focus on the "digital shadows" that they are producing: First and foremost, any capability that is being developed must be machine-facing, so that it can become part of the "capability toolbox" of that organization. With this approach, it becomes easier to use capabilities in initially unforeseen ways, because they can be easily accessed and utilized in new (digital) scenarios. On the technical level, this is mostly done through the use of Application Programming Interface (API) approaches, where every single capability in an organization must expose such an API, and ideally is even conceived and designed with that perspective in mind (this often is called an "API-First" approach).

The space of API approaches and technologies is diverse and constantly evolving. One of the reasons for that is that there is no unified global consumption platform as there is for UI Web with the browser. For this reason, the consumption side of APIs is more fragmented, and deciding on how to design APIs is not an easy task. As with any design task, everything starts with understanding the value that a product provides to a consumer, but that becomes harder when the variety of consumers becomes larger, and thus the impact of design decisions on their ease of consumption is harder to assess

- *Part 1: Understanding API Styles and Technologies:* The first part of this tutorial takes participants through the various approaches that API design can be based on nowadays. Both presenters have full-time jobs focusing on API technologies, design, and management, and thus deliver this kind of overview to various audiences on a regular basis. For the WWW2019 tutorial, the plan is to put a special emphasis on how API design for the "Open Web" looks like. In the same way as "Linked Data" and the "Web of Data" look at best practices for opening up data to an open set of consumers, the same can be done for the API space, and this will be covered as part of the first half of the tutorial.
  The UI Web thrives because it is both open and organic: providers have a lot of freedom in how they provide information and services, and in the end it is the "open market of consumers" that decides which approaches and directions ultimately gain popularity. A very similar dynamic happens in the API space: There should be a lot of autonomy for providers to decide on what they use as their design foundations, and how they implement their services. This allows

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320089

providers to make decisions that are best for the consumers that they design their services for.
But while being open and organic is essential for any system at Web scale to scale over size and time, it also helps that certain design options and decisions are not revisited from the ground up every single time. Instead, it can help a lot if there are communities that establish certain patterns, so that problems can be solved more effectively. On the UI Web, one example where this happens is the space of server-side and client-side Web app frameworks. These often have their periods of popularity, and they represent certain design approaches that were popular at that time. These frameworks come and go, but the Web itself transcends these frameworks, they were simply useful tools at a certain point in time.

- *Part 2: Supporting Continuous Change in the API Landscape:* The second part of the tutorial looks at how the individual design decisions and practices introduced in the first half can scale to scenarios with many APIs. Many organizations are facing the challenge of how to enable and empower their development teams, without centralizing too many of the design and implementation decisions. Organizations want to develop and support an "API culture" that allows them to understand and gently steer the way how APIs are designed and developed, so that ideally the combined value of all APIs and their design and development investments is maximized. We introduce the concept of "Continuous API Management" [1], which looks at APIs individually as well as at complete API landscapes. Both presenters have lots of experience in this space with regular presentations at API-focused conferences as well as with training and consulting for large organizations. The goals is to give participants a complete overview of the challenges of API management, both at the individual and at the landscape level.

REFERENCES

[1] Mehdi Medjaoui, Erik Wilde, Ronnie Mitra, and Mike Amundsen. 2018. *Continuous API Management: Making the Right Decisions in an Evolving Landscape.* O'Reilly & Associates, Sebastopol, California.

Deep Chit-Chat: Deep Learning for Chatbots

Wei Wu
Microsoft Corporation
Beijing 100080, China
wuwei@microsoft.com

Rui Yan
Institute of Computer Science and Technology
Peking University
Beijing 100871, China
ruiyan@pku.edu.cn

ABSTRACT

The tutorial is based on our long-term research on open domain conversation and rich hands-on experience on development of Microsoft XiaoIce. We will summarize the recent achievements made by both academia and industry on chatbots, and give a thorough and systematic introduction to state-of-the-art methods for open domain conversation modeling including both retrieval-based methods and generation-based methods. In addition to these, our tutorial will also cover some new trends of research of chatbots, such as how to design a reasonable evaluation metric for open domain dialogue generation, how to build conversation models with multiple modalities, and how to conduct dialogue management in open domain conversation systems.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Human-machine conversation; deep learning; chatbot.

ACM Reference Format:
Wei Wu and Rui Yan. 2019. Deep Chit-Chat: Deep Learning for Chatbots. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3320084

Starting from ELIZA in 1960s, non-goal-oriented conversational agents (a.k.a., chatbots) have never been so popular as in recent years. Take Microsoft XiaoIce as an example. The chatbot which was released by Microsoft to Chinese users first in 2014 now has attracted more than 245 million users in China, Japan, U.S., India, and Indonesia. Different from task-oriented dialogue systems, which are designed for helping people complete specific tasks such as ticket booking and restaurant reservation etc., chatbots aim to engage users in human-machine conversation in open domain for entertainment and/or emotional companionship. As a result, it is easier for chatbots to go viral among end users without specific purposes than task-oriented dialogue systems. For example, until June of 2018, users from the five countries have finished more than 30 billion conversations with XiaoIce; and on average, each conversation lasts up to 20 turns. The promising user data indicate impressive popularity of the chatbot product.

Goal of the tutorial. In contrast to the prosperity of chatbots among end users, there are few systematic introductions to approaches about how to build the conversational engines behind chatbots in research communities. The reason might be that the conventional chatbots such as ELIZA are designed with hand-crafted rules which are difficult to achieve real open domain human-machine interactions. It is only until the recent two or three years that researchers (who benefit from the large scale social conversation data available on the public Internet and the rapid advancement of deep learning approaches) begin to develop principled and data-driven approaches to build open domain conversational systems. Therefore, we believe it could be useful and valuable to give a tutorial on recent progress of chatbots, particularly on applications of deep learning approaches to building open domain dialogue engines. The community would learn the insights behind chatbots in order to fulfill the gap between task-oriented dialogue systems and non-task-oriented ones. In summary, we expect that our audience can learn the following items from the tutorial: 1) new opportunities modern chatbots, such as Microsoft XiaoIce, bring to academia, industry, and the society; 2) how to build a state-of-the-art dialogue engine with deep learning methods for chatbots; and 3) what are likely to be the future solutions for making chatbots better.

Topic and relevance. Conversational AI is catching on fire: academic conferences especially add new research tracks for conversational studies and attract unexpected growth in the number of submissions to these tracks; companies from industry are making great efforts to develop conversational products. We are entering the AI era whereby large-scale big data become more easily available and learning techniques become more powerful. We may stand at the entrance of future success in more advanced conversational systems (social chatbots and/or virtual assistants). Although this research area still faces bottlenecks and obstacles, we have witnessed a rapid surge of conversational studies in the past few years, especially the chit-chat research in open domain.

The research topic "deep learning for chit-chat" is quite relevant to the Web research community in following ways: 1) the products powered by intelligent conversational techniques can be deployed online to serve users directly, by providing chat companionship services or assisting users to complete certain tasks, and 2) the conversational techniques are deeply connected with web technologies such as content analysis, text mining, semantic matching, and search & retrieval, which are main concerns of the Web conference. Please note that there is **NO** similar proposal included into the tutorial forums of the Web conferences in recent years, and thus we believe the topic of deep chit-chat is fresh to audience of the conference.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320084

WWW'19 Tutorial on Explainable Recommendation and Search

Yongfeng Zhang
Department of Computer Science
Rutgers University
yongfeng.zhang@rutgers.edu

Jiaxin Mao
Department of Computer Science
Tsinghua University
maojiaxin@gmail.com

Qingyao Ai
College of Info. and Comp. Sciences
University of Massachusetts Amherst
aiqy@cs.tsinghua.edu.cn

ABSTRACT

Explainable recommendation and search attempt to develop models or methods that not only generate high-quality recommendation or search results, but also intuitive explanations of the results for users or system designers, where the explanations can be either post-hoc or directly come from an explainable model. Explainable recommendation and search can help to improve the system transparency, persuasiveness, trustworthiness, and effectiveness. This is even more important in personalized search and recommendation scenarios, where users would like to know why a particular product, web page, news report, or friend suggestion exists in his or her own search and recommendation lists. The tutorial focuses on the research and application of explainable recommendation and search algorithms, as well as their application in real-world systems such as search engine, e-commerce and social networks. The tutorial aims at introducing and communicating explainable recommendation and search methods to the community, as well as gathering researchers and practitioners interested in this research direction for discussions, idea communications, and research promotions.

ACM Reference Format:
Yongfeng Zhang, Jiaxin Mao, and Qingyao Ai. 2018. WWW'19 Tutorial on Explainable Recommendation and Search. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320094

1 TOPIC AND RELEVANCE

The tutorial will introduce the research and application of Explainable Recommendation and Search Systems, under the background of Explainable AI in a more general sense. Early recommendation and search systems adopted intuitive yet easily explainable models to generate recommendation and search lists, such as user-based and item-based collaborative filtering for recommendation, which provide recommendations based on similar users or items, or TF-IDF based retrieval models for search, which provide document ranking lists according to word similarity between different documents.

However, state-of-the-art recommendation and search models extensively rely on complex machine learning and latent representation models such as matrix factorization or even deep neural networks, and they work with various types of information sources such as ratings, text, images, audio or video signals. The complexity of state-of-the-art models makes search and recommendation systems blank-boxes for end users, and the lack of explainability weakens the persuasiveness and trustworthiness of the system, making explainable recommendation and search important research issues to IR, RecSys, KDD, and the Web research communities.

In a broader sense, researchers in the whole artificial intelligence community have also realized the importance of Explainable AI [3], which aims to address a wide range of AI explainability problems in deep learning, computer vision, automatic driving systems, and natural language processing tasks. As an important branch of AI research, this further highlights the importance and urgency for our IR/RecSys/KDD/Web communities to address the explainability issues of various recommendation and search systems.

Recently, a series of AI regulations have entered into force, such as the EU General Data Protection Regulation (GDPR) and The California Consumer Privacy Act of 2018, which emphasize the "principle of transparency" of intelligent algorithms, and imply the "right to explanation" of algorithmic decisions. As an important branch of AI research, this further highlights the importance and urgency for our research community to discuss and address the explainability issues of various recommendation and search systems.

Breadth. The tutorial covers two main topics, explainable recommendation and explainable search [1, 2, 4–9]. Both topics will start from very early research efforts on recommendation and search systems, such as user/item-based collaborative filtering methods, TF-IDF, and BM25 models, so that we can prepare the audiences will proper background and problem settings. The tutorial will then introduce a sequence of efforts from the community on the explainability issues of recommendation and search systems in a chronological order, so that we can help the audience to understand how the research area developed from 1980/1990s research pioneers to the most recent research advances. We will also provide a taxonomy to classify different methods for better understanding. The tutorial will cover the topics about how explainable and search techniques are applied in real-world systems such as e-commerce (Amazon, Alibaba), search engines (Google, Bing), and social networks (Facebook, Twitter).

Depth. State-of-the-art search and recommendation systems are powered by different kinds of machine learning models to estimate the relevance between queries and documents or users and items. As these models become more sophisticated, it also becomes difficult to understand how they actually work. Consequently, the lack of explainability makes it difficult for system designers to debug and improve the ranking models, or for end users to better access the search/recommendation results so as to build trust in the system. Therefore, some research efforts, albeit not necessarily under the term explainable recommendation and search, have been put to improve the explainability of recommendation and search systems.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW'19, May 13 – 17, San Francisco, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6675-5/19/05…$15.00
https://doi.org/10.1145/3308560.3320094

In terms of explainable recommendation, we will introduce research efforts on Computer Human Interaction (CHI) that attempts to understand the user behaviors in recommender systems so as to provide recommendation explanation. We will also introduce research efforts on the explainability of recommendation models, beginning from early-stage user/item-based collaborative filtering methods and content-based methods, to matrix factorization based methods, and to the recent deep learning-based approaches.

In terms of explainable search, we will introduce from both user perspective and system designer perspective. In user perspective, search system can be seen as a tool to access a huge information repository, and users should have a correct mental model of the system to know its capabilities and limitations. Based on research efforts on user behavior analysis, we will introduce how to help users understand why the search engine ranks particular documents at top positions. From system designer perspective, we will introduce how and why ranking models such as learning to rank and deep matching models output user-perceived relevances in the way that they are supposed to, based on recent advances on feature sensitive analysis and attention mechanisms.

The team. All of the presenters are active researchers in the explainable recommendation and search area. Dr. Zhang has been consistently working on Explainable Recommendation. He co-organized the first International Workshop on ExplainAble Recommendation and Search (EARS) co-located with SIGIR 2018 [10, 11]. Dr. Mao and Dr. Ai have been working on Explainable Search systems, where Dr. Mao specializes in user behavior analysis for explainable search and Dr. Ai specializes in explainable machine learning models for explainable search. The team assemble ensures that we cover both research topics, as well as research efforts from both information science and computer science communities.

2 BRIEF BIO OF ORGANIZERS

Yongfeng Zhang is an Assistant Professor in the Department of Computer Science at Rutgers University (The State University of New Jersey). His research interest is in Information Retrieval, Recommender Systems, Machine Learning, Data Mining, and Internet Economics. In the previous he was a postdoc in the Center for Intelligent Information Retrieval (CIIR) at UMass Amherst, and did his PhD and BE in Computer Science at Tsinghua University, with a BS in Economics at Peking Univeristy. He is a Siebel Scholar of the class 2015, and a Baidu Scholar of the class 2014. He has been consistently working on explainable recommendation systems ever since 2014. His recent work on the explainability of search and recommendation models include visually explainable recommendation, knowledge graph-based explainable recommendation, natural language generation for explainable recommendation, as well as explainable product search in e-commerce.

Jiaxin Mao is a postdoc from the Department of Computer Science and Technology of Tsinghua University, advised by Prof. Shaoping Ma and Prof. Yiqun Liu. He focuses on user behavior analysis of search engines and has expertise in utilizing user behavior signals to estimate their preference and satisfaction in Web Search and building click models to extract unbiased relevance feedback in different search contexts. He also served as a SIGIR student liaison for Asia region from 2017 to 2018.

Qingyao Ai is a fifth year Ph.D student advised by Prof. W. Bruce Croft in the Center for Intelligent Information Retrieval (CIIR), College of Information and Computer Sciences, University of Massachusetts Amherst. His research mainly focuses on developing intelligent retrieval systems with machine learning techniques. He actively works on applying deep learning techniques on information retrieval problems including ad-hoc retrieval, explainable product search/recommendation and learning to rank. Before his study in CIIR, he obtained his bachelor degree from Dept. Computer Science and Technology, Tsinghua University, and finished his undergraduate thesis project on click models in THUIR lab, advised by Prof. Yiqun Liu.

3 AUDIENCE

The tutorial mainly targets on IR, RecSys, Machine Learning, and Data Mining researchers and practitioners. Since we will introduce how recent NLP and Knowledge graph techniques will help explainable recommendation and search, it is also relevant to NLP, Semantic, and Knowledge Base researchers. We also introduce how explainable recommendation and search are applied in commercial real-world systems such as e-commerce, search engine, and social networks, which is relevant to industry researchers and practitioners from different areas. For prerequisite, basic understandings of information retrieval and recommendation system knowledge will be preferred, but we will introduce the basic concepts in the tutorial for better audience engagement.

4 PREVIOUS EDITIONS

This is the first edition of the Tutorial on Explainable Recommendation and Search. Before this we organized the first International Workshop on Explainable Recommendation and Search (EARS 2018) co-located with SIGIR 2018[1]. The experience and communications from the workshop helped to prepare a clear, well-organized and inspiring tutorial on this topic.

REFERENCES

[1] Qingyao Ai, Vahid Azizi, et al. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms* 11, 9 (2018), 137.
[2] Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic Explainable Recommendation based on Neural Attentive Models. *AAAI* (2019).
[3] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).
[4] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*. 165–172.
[5] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable Search Using Local Model Agnostic Interpretability. In *WSDM*. 770–773.
[6] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018. A Reinforcement Learning Framework for Explainable Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 587–596.
[7] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. Tem: Tree-enhanced embedding model for explainable recommendation. In *WWW*.
[8] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).
[9] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR*. 83–92.
[10] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2018. SIGIR 2018 Workshop on ExplainAble Recommendation and Search (EARS 2018). In *SIGIR*.
[11] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2019. Report on EARS'18: 1st International Workshop on ExplainAble Recommendation and Search. In *ACM SIGIR Forum*, Vol. 52. ACM, 125–131.

[1] https://ears2018.github.io/

Author Index

Abdollahpouri, Himan 2
Abeliuk, Andres 206
Aghakhani, Hojjat 1288
Ai, Qingyao 1330
Alaee, Sara 635
Alba, Alfredo 604, 610
Alexandridis, Georgios 335
Ali Alhosseini, Seyed 148
Almeida, Jussara 740
Almeida, Thais 975
Almeida, Virgilio 533
Alonso, Omar 344, 994, 1146, 1159, 1293
Alvari, Hamidreza 154, 1079
Amagata, Daichi 361, 398
Ambite, Jose-Luis 1240, 1327
An, Chuankai 1173
Anagnostopoulou, Evagelia 73
Anderson, Ashton 1208
Anderson, James 940
Angelidis, Iosif 840
Antelmi, Alessia 1035
Antonio Susto, Gian 1071
Antulov-Fantulin, Nino 1051
Anuyah, Oghenemaro 274, 1266
Aragon, Cecilia 799
Aragon, Harold 746
Arastuie, Makan 1039
Arcadio González, Camilo 751
Arndt, Natanael 943, 952
Aroyo, Lora 1100, 1294
Arsintescu, Bogdan 676
Ashad Kabir, Muhammad .. 981, 1180
Ashifuddin Mondal, Md 110, 1188
Attard, Judie 850, 1047, 1306
Azpiazu, Ion Madrazo 274
Azzam, Amr 237, 508
Badawy, Adam 162, 1007, 1296
Baião, Fernanda 581
Banerjee, Somnath 367
Bansal, Vidit 354
Baranova-Bolotova, Valeriya 1268
Barbosa, Simone 581
Barlacchi, Gianni 1311
Barnabò, Giorgio 484
Bar Siman Tov, Ittai 563
Bartley, Nathan 183, 206
Basu, Kaustav 7
Basu, Sugato 393
Beaton, Graeme 1131
Beck, Johannes 1051
Beigi, Ghazaleh 154, 1138, 1270
Berger-Wolf, Tanya 301
Beris, Themis 840
Berti-Equille, Laure 938
Bhamidipati, Narayan 1224
Bhargava, Preeti 1194
Bian, Shujun 404, 411
Bian, Yuanzhe 450
Bifet, Albert 1323
Bin Tareaf, Raad 148, 1297
Bizer, Christian 381
Blackburn, Jeremy 218
Blake, Darragh 904
Blanco-Justicia, Alberto 501
Blinov, Vladislav 1268
Blomqvist, Eva 1323
Bonanomi, Andrea 248
Borar, Sumit 327, 491
Bothos, Efthimios 73
Boza, Edwin 746
Bozzoli, Guillermo 262
Braganza, Samuel 746
Branson, Kim 676
Brants, Wesley 419
Braslavski, Pavel 850, 1268
Brochier, Robin 12, 283
Bron, Marc 614
Bronstein, Alexander 308
Burle, Caroline 807
C. Breuer, Pablo 1026
Calado, Pável 772
Calbimonte, Jean-Paul 1323
Calegari, Newton 807, 819
Calvo, Luz 751
Campos, Ricardo 1161
Carmel, David 326
Carvalho, Danilo 545
Castelo, Sonia 741, 975
Cattuto, Ciro 248
Caulfield, Tristan 218
Ce, Zhang 1051
Chaintreau, Augustin 569, 575
Chalkidis, Ilias 840
Chamberlain, Jon 603
Chang, Yinhong 189
Chaves Lima, Lucas 317
Chelliah, Muthusamy 1319
Chelmis, Charalampos 231
Chelmis, Charalampos 1290
Chelsy Xie, Xiaoxi 1254
Chen, Huajun 672, 678
Chen, Jiaoyan 678
Chen, Rui 285
Cheng, Lu 169, 755, 763
Christiaanse, Rob 83
Chuit, Roberto 262
Chunming 310, 459, 746, 1001
Cohen, Stephen 117, 961
Colpaert, Pieter 966
Constantopoulos, Panos 1313
Corby, Olivier 883
Costa, Rosa 581
Cremonesi, Paolo 1316
Cucchietti, Fernando 751
D. Fernández, Javier 508
Dadoun, Amine 896
Damani, Sonam 1286
Daniel, Jennifer 424, 676
Deb, Ashok 237, 850, 1007
Debruyne, Christophe 954
De Cristofaro, Emiliano 218
DeLuca, Chad 604, 610, 624
de Melo, Tiago 137, 772, 1299
Deng, Shumin 678
Destandau, Marie 17
Dhoju, Sameer 981
Diaz, Juglar 779
Dietz, Laura 672

Dietze, Stefan 137
Dikeoulias, Ioannis 1155
Ding, Ying 491, 672, 1100
Djebri, Ahmed El Amine 859
Domingo-Ferrer, Josep 501
Domingue, John 866
Domingues, Gil 1203
Dong, Yuxiao 1321
Dragovic, Nevena 1266
Duffield, Nick 285
Dumitrache, Anca 1294
Duroyon, Ludivine 988
E. Papalexakis, Evangelos 176
Ehrhart, Thibault 1232
Eißfeldt, Hinnerk 93
Elghafari, Anas 975
El Karoui, Noureddine 931, 934
Ellinger, Sarah 1194
E Robertson, Ronald 553, 959
F. Nakamura, Eduardo 742
Faloutsos, Michalis 27, 176
Farhoodi, Mojgan 1161
Farias Lóscio, Bernadette 807
Faron-Zucker, Catherine 883
Fazzone, Adriano 484
Felipe Gómez, Juan 751
Felmlee, Diane 1180
Fernanda Vélez, Diana 751
Ferrara, Emilio 162, 183, 206, 237, 1007
Ferreira Dos Santos, Eric 545, 895
Ferres, Universidad del Desarrollo, Santiago, Leo 229, 799, 1232
Fratamico, Lauren 631
Freire, Juliana 975
Freire-Vidal, Yerka 783
Fu, Yongquan 1055
Fuhrmann, Saul 1194
Fujita, Hirohisa 411
Furtado, Elizabeth 533
Furtado, Vasco 533
G. Anthopoulos, Leonidas 123
G. Beiró, Mariano 248
G. Nakamura, Fabiola 742
Galimberti, Edoardo 22
Galitsky, Boris 373
Gandon, Fabien 859, 883
Gao, Yuqing 676
Garcia, Kimberly 663, 751
García-Sáez, Artur 751
Garg, Shweta 354
Garretón, Manuela 791
Gazzaz, Samaa 1279
Ge, Jing 426
Gentile, Anna Lisa 1264
Getoor, Lise 1265
Gharibshah, Joobin 27, 176
Gharibshah, Zhabiz 176
Gimpel, Kevin 301
Giotakis, Spyros 515
Giunchiglia, Fausto 102
Goasdoué, François 988
Goel, Sharad 1296
Goel, Swati 1208
Goerzen, Matt 523
Gomer, Richard 102, 819
Gomez, Anne 1254
Gonçalves, Rafael S. 686
González, Felipe 799, 1240
Goswami, Anjan 348, 783
Graux, Damien 961
Grue Simonsen, Jakob 317, 604, 610, 994
Guille, Adrien 283
Gummadi, Krishna P. 1301
Guntuku, Sharath 461
Guo, Bin 1063
Guo, Ruocheng 169, 729
Guo, Tian 1051
Gupta, Khyatti 1286
Gupta, Somit 1288, 1299
Guzmán, José 763
H. Money, William 117
Habib, Minahz 614, 1131
Hall, Wendy 672
Han, Jizhong 729
Hansen, Casper 994, 994
Hara, Takahiro 361, 398
Hassan, Naeemul 981, 1001
Hassan, Taha 529
Heidari, Hoda 1301
Helbing, Dirk 1051
Henckell, Niklas 919
Henze, Martin 197
Herzog, Daniel 96
Higg, Joseph 450
Hodges, Jack 663
Hogan, Aidan 815
Hohlfeld, Oliver 197, 919, 1111
Hong, Liangjie 1303
Horridge, Matthew 686
Hose, Katja 714
Hosseinmardi, Homa 183
Howard, Philip 1013
Hsieh, Ping-Chun 285
Hu, Guoning 1194
Hu, Songlin 729
Huang, Di 183
Huang, Longtao 729
Huang, Roberta 1051
Huang, Tao 189
Huang, Yu 310
Hui, Pik-Mai 214
Hutchinson, Ben 1297
Ibáñez, Luis-Daniel 102
Illendula, Anurag 439
Imani, Shima 635
Inel, Oana 1294
Itti, Laurent 70
Iyer, Chander J. 1212
J. Ekaputra, Fajar 508, 1121
Jaimes, Alex 230
Jakhetiya, Vinit 461
Jampani, Dwarakanath 1121
Jannach, Dietmar 1316
Janowski, Pawel 1299
Jasser, Jasser 33
Javanmardi, Ali 1215
Jenkins, Porter 37, 294
Jiang, Hui 690
Johnson, Isaac 690, 1254
Joshi, Meghana 1286
K. Chaithanya Manam, V. 1121
Kalimeri, Kyriaki 248

Kanawati, Rushed 1306
Kandylas, Vasileios 344
Kanoria, Yash 932
Kao, Hsien-Te 183
Karlapalem, Kamalakar 705
Karras, Panagiotis 308
Kato, Linda 604, 610
Kenthapadi, Krishnaram .. 1297, 1308
Keogh, Eamonn 635, 1001, 1323
Khandekar, Sujay 450
Kiesling, Elmar 508
Kira, Beatriz 1013
Klingenberg, Brad 1098
Kohavi, Ronny 1299
Kollanyi, Bence 1013
Kostiainen, Anssi 913
Koubarakis, Manolis 840
Kreuter, Frauke 483
Kritzler, Mareike 663, 1288
Kruegel, Tina 137
Kruisselbrink, Marijn 913
Kulkarni, Ashish 354
Kumar, Ranjitha 450
Kumaraguru, Ponnurangam 130
Kumaran, Aravinda 130
Kurokawa, Mori 361, 398
Kıcıman, Emre 1297
L. Crooks, PhD, Courtney .. 755, 1026
Lalmas, Mounia 614, 1303
Lang, Carolina 262
Lange, Lukas 1159
Larson, Kate 581, 1131
Law, Edith 1131
Lees, Alyssa 1106
Lehmann, Jens 959
Leonardi, Stefano 484
Lerman, Kristina 162, 206, 1240
Leskovec, Jure 677
Li, Jason 459, 1063
Li, Mingyang 461
Li, Qiuchi 317
Li, Shuaiji 189
Liebeskind, Chaya 468
Liebeskind, Shmuel 468
Lifschitz, Sérgio 581, 655
Lim, Andrew 1131
Lima, Lanna 533, 1051
Lioma, Christina 317, 604, 610, 994
Lishchuk, Iryna 137
Liu, Feng 367
Liu, Huan 169, 1270
Liu, Tong 1111
Liu, Xi 285, 1138
López, Claudia 799
Lu, Yang 411
Luceri, Luca 237, 1007
Luiza Campos, Maria 581
Lu Jia, Adele 1055
Lyu, Yuan 361, 700
Ma, Zongyang 294
Machado, Caio 1013
Maddalena, Eddy 102
Madrazo Azpiazu, Ion 1266
Maekawa, Takuya 361, 398
Magda Florea, Adina 668, 367
Magoutas, Babis 73, 981, 1001
Malandrino, Delfina 1035
Malone, Lin 42
Mannens, Erik 966
Manolescu, Ioana 988
Mansouri, Behrooz 1161
Mao, Jiaxin 1330
María Soledad Bucalo, ... 751, 952
Masiero, Chiara 1071
Matthews, Jeanna 523, 539, 959
McGlinn, Kris 904
McGuinness, Deborah .. 672, 677, 954
Mehta, Aman 705
Mehta, Kartik 354
Mei, Qiaozhu 417
Meij, Edgar 672, 676
Meinel, Christoph 148
Menczer, Filippo 214
Menon, Abhinand 1194
Mentzas, Gregoris 73
Meta, Irene 751
Metzger, UC Santa Barbara, Miriam 974
Michahelles, Florian 663
Michel, Franck 883
Mina, An Xiao 262, 459
Miriam Carlow, Vanessa 137
Mironov, Ilya 1308
Mishra, Shaunak 1224
Mislove, Alan 553
Mitchell, Margaret 1297
Miz, Volodymyr 1188
Mohammed Abdulla, G 327
Mohapatra, Prasant 348
Mohaupt, Timon 919
Möller, Sebastian 1138
Molyneux, Sam 672
Mottin, Davide 308
Moustaka, Vaia 123
Müller, Emmanuel 308
Mulwad, Varish 672
Mumm, Olaf 137
Musen, Mark A. 686
Mustafaraj, Eni 1248
Mutharaju, Raghava 672, 1173
Naderi, Babak 1138
Najafi, Pejman 148, 975
Name, Speaker's 674, 675
Narayana, Pradyumna 393
Narayanan, Vidya 1013
Nazerzadeh, Hamid 932, 934
Ngonga Ngomo, Axel-Cyrille 840, 959
Nilizadeh, Shirin 1288
Niu, Hao 361, 398
Nyulas, Csongor I. 686
O. Talton Iii, Jerry 450
O'Hare, Gregory 655
O'Neill, Eoin 655
O'Sullivan, Declan 850, 954
Olaru, Andrei 668
Olego, Tomás 262
Oliveira, Jonice 545, 1299
Orlandi, Fabrizio 961
Oróstica, Karen 811
Özcan, Fatma 676
Ozerov, Andrey 373, 904
Palmonari, Matteo 959
Pan, Jeff Z. 678
Panchenko, Andriy 197

Papaloukas, Christos..... 840, 1311
Pappu, Aasish 1224
Parra, Denis..... 746, 763
Patricia Aires, Victoria 742
Peeters, Ralph 381, 515, 1311
Peng, Jinhua 294
Peng, Liwen..... 1055
Pennekamp, Jan..... 197, 930
Pera, Maria Soledad..... 274
Perchet, Vianney..... 934
Pérez-Solà, Cristina 501
Pertsas, Vayianos 896, 1313
Pham, Kien 975
Pirrò, Giuseppe..... 1148
Pisarevskaya, Dina 373
Poblete, Barbara 779, 811, 815
Prabhakaran, Vinodkumar 1326
Preece, Alun..... 1180
Primpeli, Anna 381, 1071
Qin, Zhiwei 189
Quadrana, Massimo 1316
R. Walker, Christopher 1026
Radtke, Norman 943
Rajan, Suju..... 931
Ramachandran, Manoharan 866
Ranganath, Suhas..... 1270
Rao, Adithya 1194
Rao, Praveen 1279
Rasiwasia, Nikhil 354, 896
Ravindran, Srinath..... 1212
Raviprakash, Nitya 1286
Razniewski, Simon..... 1155
Redfield, Olivia 1100
Reelfs, Helge 919
Rehena, Zeenat..... 110
Ren, Xiang 672
Reyes, Patricio 751, 1188
Rietz, Yannick 137, 791
Ristoski, Petar 604, 610
Rooein, Donya..... 46
Rosales-Méndez, Henry..... 50, 815
Rosati, Germán..... 262
Roselli, Drew..... 539
Rosen, Rachel..... 1100
Rosina, Alessandro 248
Roy, Deb 631
Ruback, Livia..... 545, 959
Rudnik, Charlotte..... 1232
Rula, Anisa 959
Ruswono Aryan, Peb 508
S. da Silva, Altigran 772
S. de Moura, Edleno ..772, 1039, 1161
Sagi, Tomer..... 714
Sanjay Bongale, Pratik..... 1111
Santoro, Flavia..... 581
Santos, Aécio..... 975
Sapiezynski, Piotr 553
Sarkar, Soumajyoti..... 154, 563, 1079
Sarraute, Carlos..... 262
Sawaf, Hassan..... 677
Sayyadiharikandeh, Mohsen..... 1240
Scarano, Vittorio 1035
Schaekermann, Mike..... 1131
Schler, Jonathan 563, 581, 722
Schwiegelshohn, Chris 484
Sela, Ayelet..... 563
Sen, Rijurekha..... 130
Sengamedu, Srinivasan..... 354, 419
Serra, Feliu..... 751
Seshadhri, Comandur 1317
Shaabani, Elham..... 154
Shah, Daksh 130
Shakarian, Paulo 154, 1079
Sharif, Bonita 419
Shen, Siqi 1055
Shen, Xiaoxue 1055
Sheth, Amit 439
Shinavier, Joshua..... 676
Shterionov, Dimitar..... 387
Shukla, Hemant..... 663
Sikander, Sherjeel 96
Silva, Jefferson..... 819
Silvello, Gianmaria 1071, 1311
Simperl, Elena..... 102
Singer, Alon..... 563
Singh, Shreya..... 327
Singhal, Aashay..... 705
Siolas, Giorgos..... 218, 335, 1266
Sone, Kazoo..... 393
Song, Donglei 102
Soni, Sandeep..... 206
Sonie, Omprakash..... 1319
Sorensen, Jeffrey 491
Sorici, Alexandru 668
Soursos, Panagiotis..... 840
Spasojevic, Nemanja 1194
Stafylopatis, Andreas 335, 387
Steiner, Thomas 913
Stoica, Ana-Andreea 569, 575, 1020, 1099
Stringhini, Gianluca 218
Strötgen, Jannik..... 1155, 1159
Stuckenschmidt, Heiner 1148
Sun, Sami Khenissi, Olfa Nasraoui and Patrick Shafto, Wenlong 645
Sural, Shamik..... 1319
Szlávik, Zoltán..... 1294
Tagaris, Thanos 335
Taheri, Aynaz..... 301
Talagala, Nisha 539
Tang, Jie 672, 677, 1321
Tang, Lei 404, 411
Tannier, Xavier..... 1232
Tasnim, Mayesha 961
Tavabi, Nazgol..... 206
Taylor, Jay..... 373
Teixeira, Bianca..... 581
Teixeira Lopes, Carla 1203
Tempelmeier, Nicolas..... 137
Terp, Sara-Jayne..... 1026
Tettamanzi, Andrea G.B. 859, 1232
Thain, Nithum..... 491, 1100
Thakurta, Abhradeep 1308
Third, Allan 866, 1294
Tirthapura, Srikanta..... 1317
Tochilkin, Dmitrii 393
Tommasini, Riccardo..... 56, 1323
Torres, Carmen Vaca, Johnny 255
Towsley, Don 1180
Tremblay, Serge-Eric..... 344
Troncy, Raphael 1232
Troncy, Raphaël 896
Tsirakidis, Theoklis..... 123
Tsitsulin, Anton..... 308

Tsvetkov, Yulia....1326
Tudorache, Tania....686
Turner, Liam....1180
Umarova, Khonzodakhon....1248
Ungar, Lyle....461
Urbinati, Alessandra....248
Vakali, Athena....123
Valle, Emanuele Della....1323
Vandergheynst, Pierre....1188
Van de Vyvere, Brecht....966
Vasserman, Lucy....491
Vazquez Brust, Antonio....262
Velcin, Julien....283
Venkatachalam, Akash....1111
Verborgh, Ruben....966
Verona, Letícia....581
Vidal, Maria-Esther....961
Vigna, Giovanni....1288, 1326
Vovk, Artem....393
W. Collier, Rem....655, 672
Wan, Mengting....672
Wang, Chu....404, 411
Wang, Dongsheng....317
Wang, Hanxin....398
Wang, Jianyu....310
Wang, Kuansan....677
Way, Andy....387
Weinberg, Diego....262
Weissman, Benjamin....478
Welch, Steve....310, 604, 610, 1106, 1294
Wen, Xidao....285
Whitaker, Roger....1180, 1327
Wilson, Christo....553
Wolf, Yael....714
Won Ryu, Chae....450
Wörndl, Wolfgang....96
Wu, Hua....294
Wu, Meng-Han....1121, 1329
Wu, Xintao....594
Wu, Yongning....404, 411
Wudage Chekol, Melisachew....1148, 1215
Xie, Min....367
Xie, Muhe....285
Xion, Jian....310
Xiong, Chenyan....672
Xu, Depeng....594
Xuelin Huang, Michael....69, 1329
Yan, Shen....183
Yang, Kai-Cheng....214
Yao, Mengfan....61, 231, 1290
Yonekawa, Kei....361, 398
Yousuf, Mohammad....1001
Yu, Dan....663
Yu, Yihan....799
Yu, Zhiwen....1063
Yuan, Jianping....1063
Yuan, Shuhan....594
Z. Pan, Jeff....672
Zafarani, Reza....1292
Zaim, Mariam....1121
Zannettou, Savvas....218, 553
Zeni, Mattia....102
Zequeira Jiménez, Rafael....1138
Zhai, Chengxiang....348
Zhang, Da....404, 411
Zhang, Fanfang....189, 672
Zhang, Ningyu....678
Zhang, Qiuyun....1063
Zhang, Wei....676
Zhang, Wen....678
Zhang, Yongfeng....1330
Zhang, Zuohua....404, 411
Zhou, Ke....614
Zhou, Xinyi....1292
Zhou, Yan....729
Zia, Leila....1208
Zikos, Nikos....123
Zois, Daphney-Stavroula....1290

OUR PARTNERS AND SPONSORS

Diamond

Gold

Silver

Bronze

Others

The Web Conference 2019 was organized by Web4Good, a California 501(c)(3) non-profit organization.